Micro-Facial Movement Detection Using Spatio-Temporal Features

by

Adrian Keith Davison

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

Supervised by Dr. Moi Hoon Yap, Mr. Cliff Lansley, Dr. Nicholas Costen and Dr. Kevin Tan

Faculty of Science and Engineering
School of Computing, Mathematics and Digital Technology

MANCHESTER METROPOLITAN UNIVERSITY

February 2016

Abstract

Micro-facial expressions are fast, subtle movements of facial muscles that occur when someone is attempting to conceal their true emotion. Detecting these movements for a human is difficult, as the movement could appear and disappear within half of a second. Recently, research into detecting micro-facial movements using computer vision and other techniques has emerged with the aim of outperforming a human. The motivation behind a lot of this research is the potential applications in security, healthcare and emotional-based training. The research has also introduced some ethical concerns on whether it is okay to detect micro-movements when people do not know they are showing them.

The main aim of this thesis is to investigate and develop novel ways of detecting micro-facial movements using features based in the spatial and temporal domains. The contributions towards this aim are: an extended feature descriptor to describe micro-facial movement namely Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) combined with Gaussian Derivatives (GD); a dataset of spontaneously induced micro-facial movements, namely Spontaneous Activity of Micro-Movements (SAMM); an individualised baseline method for micro-movement detection that forms an Adaptive Baseline Threshold (ABT); Facial Action Coding System (FACS)-based regions are proposed to focus on the local movement of relevant facial areas.

The LBP-TOP with GD feature was developed to improve on an established feature and use the GD to enhance the facial features. Using machine learning, the method performs well achieving an accuracy of 92.6%. Next a new dataset, SAMM, was introduced that improved on the limitations of previous sets, including a wider demographic, increased resolution and comprehensively FACS coded. An individualised baseline method was the introduced and tested using the new dataset. Using feature difference instead of machine learning, the performance increased with a recall of 0.8429 on the maximum thresholding and a further increase of the recall to 0.9125 when using the ABT. To increase the relevance of what is being processed on the face, FACS-based regions were created. By focusing on local regions and individualised baselines, this method outperformed similar state-of-the-art with an Area Under Curve (AUC) of 0.7513.

The research into detecting micro-movements is still in it's infancy, and much more can be done to advance this field. While machine learning can find patterns in normal facial expressions, it is the feature difference methods that perform the best when detecting the subtle changes of the face. By using this and comparing the movement against a person's baseline, the micro-movements can finally be accurately detected.

Acknowledgements

Throughout my research and preparation of this thesis, many people have guided my understanding and knowledge that allowed me to form the work presented. Firstly, I would like to thank my Director of Studies, Dr. Moi Hoon Yap, for her kindness, patience and overall confidence in my abilities from the very beginning. I also want to express my utmost gratitude to my other supervisors Mr. Cliff Lansley, Dr. Nicholas Costen and Dr. Kevin Tan, who all provided their own unique perspective in different aspects of my work and contributing greatly when I was in need.

I have received help and advice, directly or indirectly, from my peers and colleagues throughout my study. These people motivated me to share ideas and build relationships to further expand my knowledge. From the Emotional Intelligence Academy, Mr. Harry Lansley and Mr. Jordan Lansley. From MMU, Dr. Daniel Leightley, Dr. Choon Ching Ng, Mr. Brett Hewitt, Mr. Ezak Fadzrin, Mr. Connah Kendrick, Ms. Gemma Stringer, Mr. Nadim Baharum and everyone who kindly agreed to participate in my emotional inducement experiment to form the SAMM dataset. I would also like to thank all the FACS coders who contributed their efforts to code the ground truth for this dataset.

In my first year of study, the administration on my study at MMU was overlooked by Ms. Rita Kenny, who I thank for always being attentive a busy role. Later, the research administration roles were taken over by Ms. Megan Schofield, Ms. Francesca Robinson and Ms. Kristina Ganchenko. As well as always being available to help with a multitude of requests, they were key to helping with the success of the Science and Engineering Research Symposium 2015. Thanks should also go to Dr. Mike Dempsey and Dr. Sam Illingworth for guidance with leading this event.

I would like to thank my friends Ms. Laura Bennett, Ms. Abbie Bywater and Miss. Bethany Woodcock for always being there whenever I needed advice through the high and low points. I would also like to thank Ms. Beth Scott for always

supporting me no matter what, and being one of few people to truly understand when I was in need.

My most sincere thanks goes to my family, for always believing I could succeed, even when I did not. To my Grandparents, Mr. Keith Butterworth and Mrs. Hazel Butterworth, who have always been available if I need them, and always help me to think positively. To my Mother, Ms. Julie Davison, I am unable to express how much I want to thank her for always listening and defending me. For always making me laugh, through the difficult times we both have had in the past. Finally, I would like to thank my Father, Mr. Garry Davison, who passed away before he could see me begin this journey, for inspiring me to focus my interest in technology.

I am grateful for MMU for providing the studentship for me to complete this thesis, and provide valuable experience in teaching at a university level.

Contents

List of Figures

List of Tables

List of Abbreviations

2D 2-Dimensional

3D 3-Dimensional

AAM Active Appearance Model

ABT Adaptive Baseline Threshold

ASM Active Shape Model

AU Action Unit

AUC Area Under Curve

CASME Chinese Academy of Science Micro-Expressions

df Degrees of Freedom

DFT Discrete Fourier Transform

DPM Deformable Part Models

EMFACS Emotion FACS

FACS Facial Action Coding System

fps Frames per Second

FP False Positive

FPR False Positive Rate

FN False Negative

GD Gaussian Derivatives

HOG Histogram of Oriented Gradients

3D HOG 3D Histogram of Oriented Gradients

HOOF Histogram of Oriented Optical Flow

JACFEE Japanese and Caucasian Facial Expressions of Emotion

LBP Local Binary Patterns

LBP-TOP Local Binary Patterns on Three Orthogonal Planes

LED Light Emitting Diode

MDMO Main Directional Mean Optical-flow

METT Micro-Expression Training Tool

MCC Matthew's Correlation Coefficient

OOB out of bag

PWA Piecewise Affine

SAM Self-Assessment Manikins

SAMM Spontaneous Activity of Micro-Movements

SAT Summed Area Table

SD Standard Deviation

SMIC Spontaneous Micro-Expression Corpus

SVM Support Vector Machines

TIM Temporal Interpolation Model

TP True Positive

TPR True Positive Rate

TN True Negative

RF Random Forests

ROC Receiver Operating Characteristic

ROI Regions of Interest

USF-HD University of South Florida-High Definition]

VBM3D Video Block Matching and 3D filtering

YorkDDT York Deception Detection Test

List of Publications

This thesis is based on material from the following publications:

1. Adrian K. Davison, Moi Hoon Yap, Nicholas Costen, Kevin Tan, Cliff Lansley, Daniel Leightley (2014), "Micro-facial Movements: An Investigation on Spatio-Temporal Descriptors", Computer Vision - ECCV 2014 Workshops, Lecture Notes in Computer Science Volume 8926, 2015, pp 111-123.

2. Adrian K. Davison, Moi Hoon Yap, Cliff Lansley (2015), "Micro-Facial Movement Detection Using Individualised Baselines and Histogram-Based Descriptors", Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on, 2015, pp 1864-1869.

3. Adrian K. Davison, Moi Hoon Yap, Nicholas Costen, Cliff Lansley, Kevin Tan (2016), "SAMM: A Spontaneous Micro-Facial Movement Dataset", IEEE Transactions on Affective Computing. Status: Accepted.

Dedicated to my Parents

Chapter 1

Introduction

This Chapter introduces the research and terms described in the analysis and detection of micro-facial movements. The main terminology used throughout this thesis is defined alongside the problem statement, thesis contributions and the thesis structure.

1.1 Background

Facial image analysis is a well established research area that has produced many novel works in areas such as face detection, facial expression analysis, gender identification, face age estimation and blink recognition [1–4]. Micro-facial movement research is relatively new [5–9] when compared with facial expression analysis. Although loosely related due to the facial expression aspect, these two topics should be looked upon as different research problems. The main example of why this is true is facial expressions tend to be large and distinct, whereas micro-facial movements are very quick and subtle muscle movements. Throughout this thesis, many terms are used from previous literature and some new terms to help explain this work are defined. The term *micro-facial expressions* (and some similar variations) is used in most current work to globally describe the movements from its origins in psychology [10]. Although this term is used in this thesis, it specifically is meant to describe the movement when being related to emotions. For example, if a method uses machine learning to classify the micro-expressions into distinct classes. To follow a more objective approach, the terms *micro-facial movements* or *micro-movements* are used interchangeably to describe only the movement of

the face, with no emotional interpretation. This approach treats all movements as their basic muscle activations to avoid assumptions and bias where possible.

The research interest for this work is detecting micro-facial movements. Current literature has focused on two interpretations of this problem, the first being micro-expression recognition, where machine learning is trained using (limited) micro-expression datasets and then tested to see how well the micro-expression are assigned to emotion-based categories. These usually form simple classes such as positive and negative, or into specific emotion categories like happiness and surprise. The second method takes a more objective approach by not using machine learning to determine specific micro-movement patterns. Instead the difference changes in features is calculated, where an increased magnitude of change indicates a movement. The larger the magnitude, the bigger the movement will be. This approach usually forms a binary classification of movement and non-movement, with non-movement being equivalent to the person not activating any facial muscles. This method does not assume any emotional link with the movements, limiting the possibility of automatic interpretation. However, in a real-world environment, people perform micro-movements in so many different ways that modelling each into distinct classes would be unrealistic. Starting with a detection foundation would be beneficial to provide an initial point from which further interpretation can follow.

Although the potential for micro-movement analysis in computer vision is huge, core aspects of development need to be greatly improved to get accuracy rates to match facial expression analysis. Challenges facing this field can include the lack of available datasets, the speed and subtlety of micro-movements and the noise that occurs through digital capture devices, especially through high-speed cameras that are used to capture more information for analysis. Other points are extremely difficult to control, such as head movements being interpreted as micro-movements, consistent lighting and a wide enough demographic of participants in datasets. Finally, all datasets currently available induced participants in a laboratory, as there is currently no ethical procedure to induce micro-movements outside of these controlled conditions.

The current most popular method is to recognise micro-facial expressions using machine learning algorithms. Simple feature descriptors using histogram of oriented gradients [11] and clustering is used to represent Action Unit (AU)

categories, however there is limited data used (13 movements) and posed micro-expressions are used. Other methods use classification learning algorithms, mainly Support Vector Machines (SVM) [12], by extracting features of the face and training based on the output. Another common processing method is to split the face into blocks or regions [6, 8, 9], where local feature are the prime source in extracting the features rather than using the whole face. Global representations such as these are less suitable for subtle movements occurring in a small portion of the face.

This thesis investigates methods to detect micro-facial movements, in an objective manner, with the potential to be applied into real-world problems. As humans find it difficult to spot micro-movements consistently [13], using computer vision can increase user's ability to spot through training, or as an aid during situations where a human would struggle to multi-task. Already in security the research behind these hidden signal have been used [14], so applying a system to automatically find these movements would be highly beneficial. Even though most recent application possibilities have been aimed towards security, other areas include studying and improving the lives of people with health problems. Depression can be very debilitating to the point of suicide, and many people who have the illness do not, or cannot, share their feelings. Finding hidden emotions that are involuntarily shown can improve the understanding of health professionals in each individual case. Similarly, a system that detects these movements can train others in how to spot more effectively, and how to better understand people from an emotional point of view.

One of the biggest challenges facing micro-movement detection is the lack of available datasets for validating the algorithms. Current datasets [15–17] have limitations such as a limited demographic (i.e. only using one ethnicity), not many movement examples compared to the facial expression counterpart, and the difficulty to induce micro-facial movements naturally and spontaneously. The final point is the most difficult to solve. In a real-world environment, micro-facial movements occur when someone attempts to hide their true emotion. Often a high-stakes scenario is required to induce effectively, and only particular triggers for each individual will induce a genuine response. All datasets have been created in a laboratory scenario, where stimuli (usually video clips) are shown to participants and they are told to only show a neutral face (i.e. do not show a facial expression). Immediately the environment is not natural, and can affect how the

participant reacts to the stimulus. Ideally, people should be filmed without them knowing, and inducement attempted without them realising. Unfortunately, this is constrained by ethical issues on recording people without their knowledge and trying to psychologically manipulate emotions to how you want. Even if this could be done, micro-movements are not guaranteed without the person knowing they have to hide their true feelings. Further, certain emotions, such as anger, are very hard to ethically induce as the consequences of genuinely angering someone could be disastrous.

This Chapter is organised as follows: the current Section and Section 1.2 provide the background and motivation behind this work; Section 1.3 describes the issues surrounding the work and provides a overall problem statement; Section 1.4 states the aim and objectives of this work; Section 1.5 lists the contributions made by this thesis; finally, Section 1.6 summarises the structure of the thesis.

1.2 Motivation

The fast-growing research area of micro-facial expression analysis is largely driven by the potential of deploying these algorithms to real-world environments. Research behind the psychology of micro-expressions has focused on the ability of humans to detect them and then decipher what and why a person felt they needed to try and hide their true feelings. However, humans still struggle to detect micro-expressions in real-world situations where being able to see, remember and interpret all micro-expressions would be challenging.

Using a system that can detect micro-expressions or movements would aid the user in reviewing detected micro-movements and help with interpretation. Unfortunately, even though a computer assisted system would be better at processing movements, it takes more steps to differentiate between noise and a real micro-movement. Many recent systems try to recognise using machine learning or by using computationally expensive image processing techniques, both of which would not be suitable for applying the system to a real-world environment.

The following points summarise the motivation behind creating an accurate micro-movement detection algorithm:

1. These quick movements have been found to occur when someone is trying to hide their true emotion [10, 18]. When a person attempts to hide their feelings, showing micro-movements cannot be stopped in many cases. Being able to extract this information would be useful for applications in security and deception detection [13, 19, 20].

2. It has the potential to help people with health problems such as depression [21] and neurological disorders such as facial paresis [22]. These applications aim to help others who cannot readily identify the problems they are facing day to day.

3. For training humans to understand micro-expressions better by combining technology and previous techniques to allow for improved integration of cross-discipline research [23–25].

1.3 Problem Statement

The nature of an emerging field means that research is limited and tends to be exploratory rather than focused on already established work. In Chapter 2, current methods and algorithms for recognising, classifying and detecting micro-movements are discussed. Many different feature descriptors are used to describe micro-movements in the spatial or temporal domain, and ways of determining if a movement has occurred include machine learning and feature difference analysis.

The main problem faced in the analysis of micro-movements, is how to detect a movement that is extremely subtle and fast moving. Many obstacles occur during processing, such as small head movements being mistaken for real micro-movements and the speed being difficult to capture with a typical frame rate of 30 Frames per Second (fps). Further, the amount of datasets, especially compared with normal facial expressions, is very limited due to the difficulty of spontaneously inducing micro-expressions in a laboratory environment. The lack of a standardised emotional inducement protocol is also a limitation of being able to create these datasets easily.

> *The logical progression of developing methods of detecting micro-facial expressions or movements is to use techniques already established in facial analysis. Unfortunately, different feature representations have not*

produced similar results for micro-analysis compared with normal facial expression analysis. Machine learning techniques also struggle, as they look for a pattern that exists. Larger facial movements may be quite distinct in their representation in feature space, but the sparsity of micro-movements reduces the classification accuracy of even well-established learning algorithms. A way of objectifying the micro-movements of the face is required, where they are treated as basic facial muscle movements, with further interpretation, done by a user afterwards. This thesis investigates machine learning methods common in current research. Moreover, novel techniques of representing micro-movements are discussed with detection completed by taking into account a person's individual expression baseline.

1.4 Objectives

The primary aim of this research is to test and evaluate established feature extraction methods, learning algorithms and feature difference methods on high-speed video to develop a novel way of detecting micro-facial movements. This will help humans find and interpret emotion based on these movements. To achieve this aim, the following objectives have been established:

1. To investigate the potential of being able to recognise micro-movements from non-movements using temporal features and machine learning algorithms.
2. To create a new spontaneous micro-movement dataset by conducting an emotional inducement experiment.
3. To explore and compare different feature descriptors that best represent micro-movements for accurate detection.
4. To propose an objective method of detecting micro-facial movements using localised features.
5. To evaluate the performance of the proposed methods against benchmark algorithms and datasets.

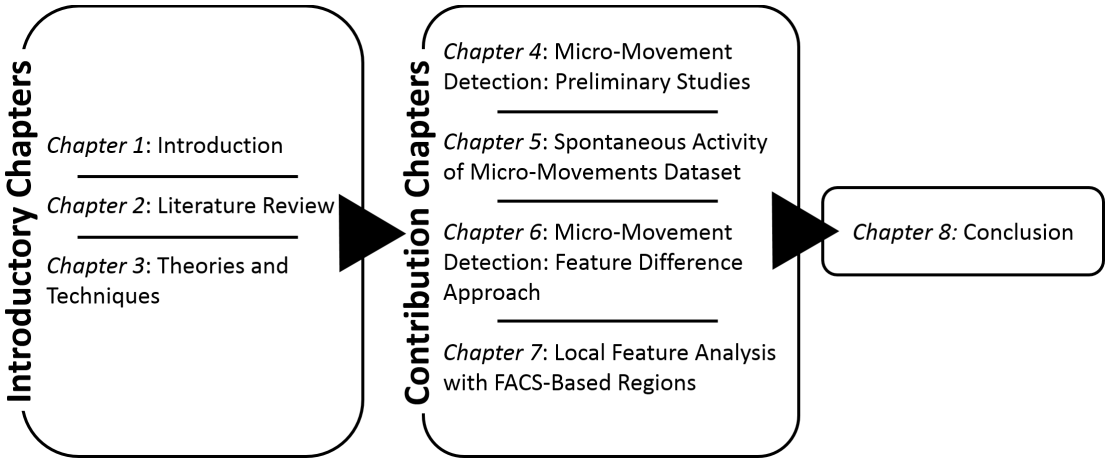

FIGURE 1.1: The organisational structure for this thesis. The first three Chapters introduce the research to be presented as well as a thorough literature review on previous work and techniques. The next four Chapters are the contributions made in this thesis and the final Chapter concludes this work.

1.5 Contributions

The main contributions of this thesis are as follows:

1. An extended feature representation using Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) combined with Gaussian Derivatives (GD) is created for micro-movement recognition
2. A new dataset (SAMM) is collected and FACS-coded using a new design of emotional inducement protocol
3. An individualised baseline micro-movement detection method using Histogram of Oriented Gradients (HOG) features and a temporal difference method is proposed
4. A novel micro-movement detection method based on FACS-based regions and individualised baselines with 3D HOG features is introduced.
5. Evaluation of the proposed methods and benchmark algorithms on SAMM and Chinese Academy of Science Micro-Expressions (CASME) II datasets.

1.6 Thesis Organisation

As shown in Fig. 1.1, this thesis is split into two main sections: introductory chapters and contribution chapters. The first section consists of three Chapters, the first of which is the current Chapter introducing the work presented in the thesis and outlining what to expect from the research.

Chapter 2 presents fundamental knowledge and a review of the literature relating to micro-movement detection. Given the nature of the field, some psychological research is included to form a foundation on which a micro-movement detection system should be based.

Chapter 3 provide the technical information on the techniques explored for micro-movement detection. This includes feature extraction methods, machine learning approaches, feature difference methods and automatic detection of micro-movements using defined thresholds.

The second section includes four contribution chapters. Chapter 4 investigates the recognition of micro-movements using a temporal feature descriptor named LBP-TOP with GD. The classification is then completed using SVM [12] and Random Forests (RF) [26] to test two different machine learning methods.

Chapter 5 introduces a new spontaneous micro-movement dataset created to address the limitations of current benchmark micro-movement datasets [15–17]. Further, emotional inducement protocols were designed with replication in mind and the data would allow for the advancement of research into micro-facial movements in computer vision and the psychological communities.

Chapter 6 moves towards the more objective micro-movement detection by using a feature difference method, the process would treat the movements as muscle activations and model peaks when the magnitude of temporal frame differences were analysed. Further, an ABT is introduced based on the non-movement sequence.

The last of the contribution chapters, Chapter 7, presents newly developed local FACS-based regions that focus on areas of the face relevant to facial muscle movements. Each of the 26 regions correspond to a different part of the face, meaning all local changes in that area can be described. Micro-movements can be detected using individualised baselines, feature difference on each region to localise

important changes and finally automatic peak detection to find the onset, apex and offset. To test the robustness of this method, it is validated on the two most recent micro-movement datasets: SAMM and CASME II.

Finally, Chapter 8 concludes this thesis with a summary of contributions, the limitations faced in the field of micro-movement analysis and the future research direction.

Chapter 2

Literature Review

This Chapter presents a review to the background of micro-facial movements and the current state-of-the-art methods in recognising and detecting the movements using computer vision and machine learning. The current benchmark micro-movement datasets are also explored, focusing on how they contribute to the field and the limitations.

2.1 Introduction

In this chapter, an overview of the current literature surrounding emotion, facial expressions and micro-facial expressions is presented. There is a large amount of study around the psychology of emotion, dating back to 1872, and the discussion around this is not exhaustive but allows for the context behind the contributions of this thesis. A more in depth analysis of this area can be found in various psychology based papers [10, 27–29]. To fully understand the complexity of facial muscles, a brief overview of face physiology is also discussed.

Facial expressions, or movements that we typically do on the face to express emotion, pain and other factors, can be analysed using a computer quite easily due to there large distinct movements. An introduction to some literature in this field is discussed to highlight the differences between this field and micro-movement analysis. The focus of this thesis is to study and develop novel ways of detecting micro-facial movements on the face, attempting to overcome the challenges faced by the current state of the art method described.

In contrast to facial expression datasets such as the Cohn-Kanade original and extended versions [30, 31] and the MMI dataset [32], micro-movement datasets are quite scarce and only a few are publicly available for research purposes. The current datasets that have been used in research, whether they are available or not, are described to highlight the large differences between datasets and what are the limitations of the ones in use.

Finally, some real-world problems are discussed with the psychological study of cultures and how they suppress their emotions in different ways and how this can relate to micro-movements. The applications where this research has not yet fully integrated, due to being in it's infancy, but would greatly benefit organisations involved with security, health and animation. The potential areas and application types are discussed

An outline of this Chapter is as follows: Section 2.2 introduces the physiology of the facial muscles and how activation of these muscle create facial expressions. The study of facial expressions are also discussed. In Section 2.3 the universal emotion theory is summarised and why it is important for informing computer vision algorithms. Section 2.6 and 2.7 discusses current methods of facial expression and micro-facial expression analysis respectively. The benchmark datasets containing micro-facial expression videos are described in Section 2.8. Real-world applications of micro-facial expression detection are discussed in Section 2.9, which focuses on security, deception detection and the development of emotional intelligence. The final sections focus on the research direction and a summary of this chapter.

2.2 Physiology of the Face and Facial Expressions

Facial expressions appear from neurological signals produced by the brain to activate facial muscles and cause the skin to be distorted into what is commonly accepted as a meaningful expression of emotion [34].

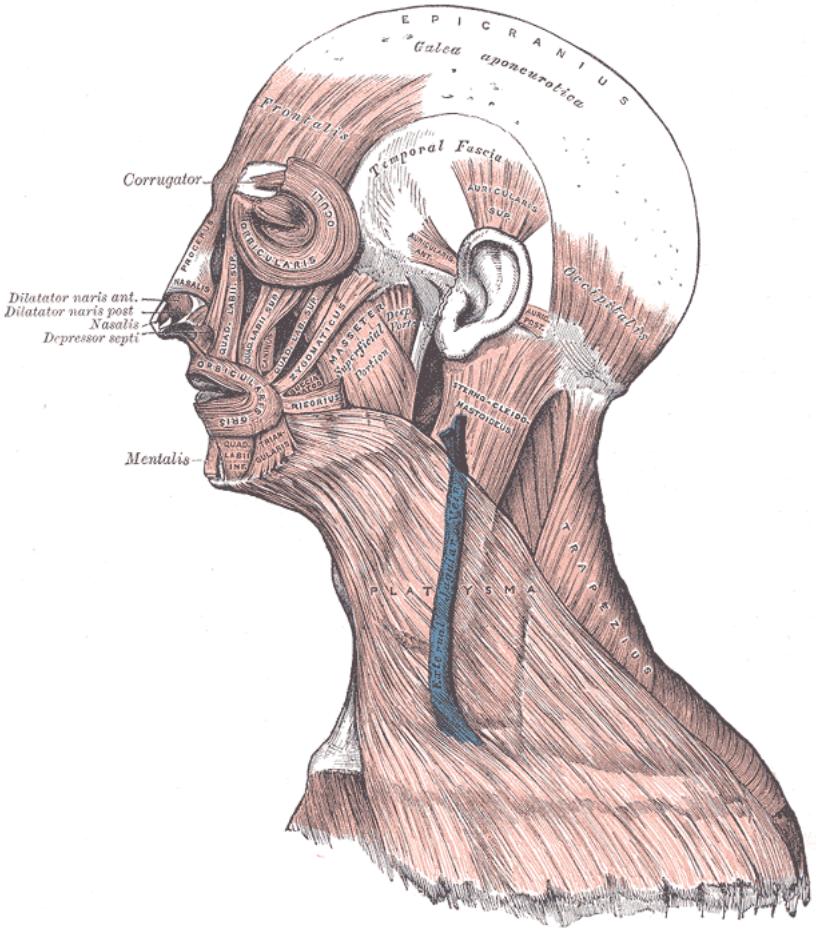

FIGURE 2.1: Each facial muscle is attached to the skull and contract towards the bone it is fixed to. The profile is used here as the facial muscles are 'mirrored' on the other side of the face. This diagram is reproduced from Gray's analysis of anatomy [33].

2.2.1 Face Anatomy

The face itself comprises of a complex system of skeletal muscles that contract and pull the skin of the face to deform it from a relaxed state. As can be seen in Fig. 2.1, the many different muscles have particular roles to move parts of the face. For example, the obicularis oculi is a circular shaped muscle around the eye socket that makes the eyes squint. The zygomaticus major pulls the corners of the mouth upwards and is activated when a person smiles.

Facial muscles are made up of fast moving fibres that can contract and relax in less than 20 ms including a latency period where the muscle has to receive instruction from the central nervous system [35].

2.2.2 The Human Visual System

Human visual perception depends on many factors such as light intensity, distance from and size of the object. The human visual system can form, transmit and analyse 10-12 images per second [36]. Light intensity pooling lasts around 15 ms and samples images from the retina as spikes in activity. This suggests that rapid movements,such as 1/25th of a second, are on the limits of human perception, and can be difficult to interpret if seen.

2.3 Universality of Emotion

Facial expression research accelerated through the 1970s, and this modern theory on 'basic' emotions has generated more research than any other in the psychology of emotion [37].

Facial expressions have strong scientific evidence suggesting they are universal rather than culturally defined or learned. Charles Darwin first proposed that there were a specific set of expressions that he had observed and studied that were universal [29]. Later, Tomkins [38] expanded on Darwin's studies with Affect Theory, emphasising the importance of the face conceiving emotion in terms of eight discrete, quite different affects, rather than two or three affective dimensions.

These studies helped Ekman [10] first conclude facial expression were innate with experiments in Papa New Guinea, where local tribes, who had very little contact with outsiders, especially Westernised culture. These pre-literate people were able to recognise the facial expressions and related emotions in images created by Darwin and Tomkins.

During his time with a tribesman, Ekman asked the man to show what his face would look like if:

- Friends had come
- His child had just died
- He was about to fight
- He stepped on a smelly dead pig

TABLE 2.1: The seven universal emotions and universal triggers [10].

Emotion	**Universal Trigger**
Happiness	Pleasure
Sadness	Loss of a valued object or person
Anger	Interference with goals
Fear	The threat of harm
Surprise	A sudden and unexpected event
Disgust	Offensive in nature
Contempt	Immoral action

These questions corresponded to a particular emotion that was expected as a response. In order of the question list, the emotions were happiness, sadness, anger and disgust. Posed examples of the facial expressions given by a Papa New Guinea tribesman, in order from left to right, can be seen in the top row of Fig. 2.2. Due to image copyright, the original images of the tribesman are available in [10].

When an emotional episode is triggered, there is an impulse that cannot be controlled which may induce one of the 7 universal facial expressions (happy, sad, anger, fear, surprise, disgust or contempt). Table 2.1 describes each of the seven emotions and the universal trigger for each emotion and Fig. 2.2 shows posed facial expressions of these emotions. While the trigger for the emotion is fixed, how people are triggered will vary person to person. For example, someone may be fearful of heights whereas another find pleasure in bungee jumping.

2.3.1 The Facial Action Coding System

The Facial Action Coding System (FACS) was first published by Ekman and Friesen [39] as a research tool to measure any facial expression that a human can perform. It was designed to objectively understand the facial muscle movements with no inference to emotion i.e., how muscular action is related to facial appearances.

Each observable component of facial movement is called an Action Unit (AU) and all facial expressions can be broken down into their constituent AUs. FACS describes the criteria for observing and coding each AU, along with a description of how the AUs appear in combinations [39, 40]. Even though there is specific

FIGURE 2.2: The 7 universal facial expressions defined by Ekman [10, 28]. Top, left to right: happy, sad, anger and disgust. Bottom, left to right: fear, surprise and contempt.

descriptions on how the movements are performed, the exact appearance changes will change from one person to another. For example, bone structure, facial muscles, fat deposits and permanent wrinkles will all change how a movement looks visibly to another person or computer system. However, the underlying muscle structures remain the same.

The AUs in the FACS manual are presented in two main groups: upper face and lower face actions. Each main group is then split into sub-groups. A breakdown of the groupings and AUs belonging to that group is shown in Table 2.2. When training to become a FACS coder, a person would usually begin with the upper face, and then move on to the lower face actions. There are also head and eye movement descriptor codes, however these have not been included for simplicity.

As with all muscles in the human body, the facial muscle can be contracted in varying strengths. In FACS, these variations are classed as intensity scores, and are termed A, B, C, D and E to refer to how strong the action is. A is barely detectable or the lowest intensity, and E is the maximum strength possible and therefore the highest intensity. For the work in this thesis, the intensity scoring is rarely used, however in later Chapters, the temporal phase is described. This refers to the onset, apex and offset of a movement, with the apex being the highest intensity during the movement's temporal sequence.

TABLE 2.2: FACS AUs defined as two main groups and split into sub-groups based on their location and movement style.

Main FACS Groups	Sub-Groups	AUs
Upper Face	Eyebrows	4
	Forehead	1, 2
	Eyelids	5, 6, 7, 43, 45, 46
Lower Face	Up/Down	9, 10, 15, 16, 17
	Horizontal	14, 20
	Oblique	11, 12, 13
	Orbital	18, 22, 23, 24, 28
	Miscellaneous	8+25, 19, 21, 29. 30, 31, 32, 38, 39

2.4 Micro-Facial Expressions

When a person consciously realises that a facial expression is occurring, the person may try to suppress the facial expression because showing the emotion may not be appropriate or could be due to a cultural display rule [23]. Once the suppression has occurred, the person will mask over the original FE and cause a transient facial change referred to as a micro-facial expression [10, 41]. In a high-stakes environment, micro-facial expressions tend to become more likely as there is more risk to showing the emotion.

The duration of a micro-expression is very short and is considered the main feature that distinguishes them from a facial expression [42], with the general standard being a duration of no more than 500 ms [43]. Other definitions of speed that have been studied show micro-expressions to last less than 250 ms [18], less than 330 ms [44] and less than half a second [13]. Following Ekman and Friesen as first to define a micro-expression [45], a usual duration considered is less than 200 ms.

Experiments by Matsumoto and Hwang [46] summarise a micro-expression to be less than half a second with these experiments looking into whether training humans in detecting micro-facial expressions was effective. The findings showed that training improved the ability of reading micro-expressions and the skill was retained a few weeks after the initial training. Training humans can be time consuming and expensive, so looking into ways of aiding a person when detecting subtle movements would make training more accessible.

For humans, detecting micro-facial expressions can be difficult and usually requires a lot of specialist training. The spotting accuracy of humans peaks around 40% [13], and so analysis using computer algorithms incorporating machine learning and computer vision can be employed to aid in the detection of micro-facial expressions.

2.5 Cultural Influences

A topic of debate in Psychological literature is Emotion regulation, which has grown in popularity since the 1990s. Emotional suppression across cultures has been comprehensively analysed [23, 47–52]. In East Asian cultures, emotional suppression is encouraged as value is placed on adjustment of behaviour to the interpersonal context. In contrast, Western cultures do not encourage suppression and value autonomy and the expression of one's true abilities [24].

Biehl et al. [53] found that there were cultural differences between different nationalities when recognising emotions and rating the intensities. Subjects from Hungary, Japan, Poland, Sumatra, United States, and Vietnam viewed an image set named the Japanese and Caucasian Facial Expressions of Emotion (JACFEE), which contained emotional expressions of an equal number of people who were Caucasian and Japanese. A high agreement was obtained between participants on what emotion they were looking at, however difference across nations were found for the photos of anger, contempt, disgust, fear, sadness, and surprise. A similar results was found for the intensity ratings. A study into how well people from different nations could spot micro-facial movements from varying countries would help determine if there are cross-cultural differences in micro-movements, or if they are similarly universal.

2.6 Analysing Facial Expressions with Computer Vision

Understanding context and how powerful human emotions are, is fundamental to developing an effective detection system using a computer. Studies into people posing facial expressions, under the self-perception theory in Psychology, have

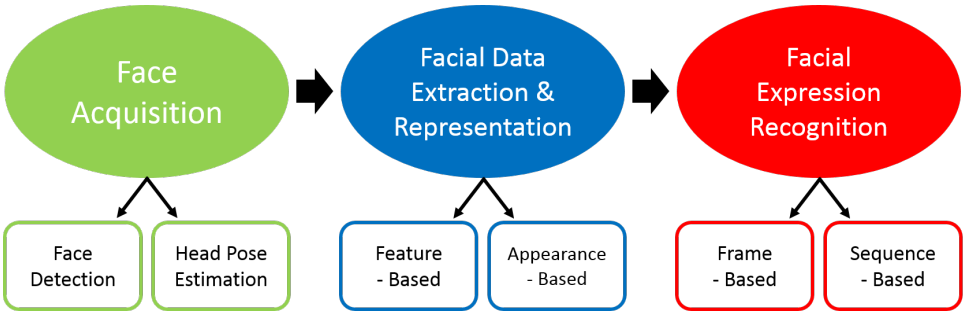

FIGURE 2.3: The general pipeline of a facial expression analysis system. Image reproduced from Tian et al. [56].

found that regions of the brain that associated with enjoyment activate when a person voluntarily smiles [54] and more recently experiments with a large number of participants (170) [55] found that voluntarily smiling under stressful situations helped reduce heart rate compared with participants who kept a neutral expression.

2.6.1 Facial Expression Analysis

Automatic facial expression analysis typically refers to a computer system that can take visual information and automatically analyse the facial motions and features to recognise particular facial expressions. It is an active research area and has applications in human-computer interaction, computer generated avatars and emotion classification [2]. Due to the rich amount of research in this area, it is impossible to discuss every method, and so important developments and techniques are discussed. More in-depth analysis can be found in [2, 56, 57].

The pipeline of facial expression analysis can be generalised into three stages: face acquisition, where the face is detected and preprocessed; facial data extraction, where the facial expression images are typically processed using a feature-based or appearance-based method; and recognition, where the expression is classified by machine learning. Fig. 2.3 summarises this pipeline.

Any facial expression analysis system needs to be able to analyse facial actions regardless of a person's gender, age, ethnicity or culture. Humans with normal facial muscle structures exhibit the same movements, so modelling a general movement pattern is possible. For micro-movement analysis, discussed in Section 2.7, the same requirement of person independent analysis is set.

2.6.2 Face Acquisition and Preprocessing

Firstly the face needs to be identified using a face detector. Viola and Jones [58] developed a robust real-time face detector based on a set of rectangle Haar features, which was later expanded by Lienhart et al. [59] to include more features for better performance. Many face detectors could only detect frontal faces, as these were the features they were trained on. For most facial expression problems, frontal face detectors are adequate, however for face detection 'in the wild', detectors need to be robust to changing head pose [60, 61].

After detecting the face preprocessing is applied and can include aligning faces, normalisation, converting images to grey-scale, histogram equalisation (to even lighting) and masks that can remove unnecessary pixel data such as hair or ears.

To address spatial differences of detected face regions, faces should be normalised by using alignment techniques. Alignment can be done by manually marking feature points, such as the eyes, on faces and aligning all faces to these feature points. Doing this can be very time-consuming and may be inaccurately aligned. Using this fully supervised method means that points must carefully be chosen based on the specific object class, and then select many example image patches of these points. Therefore, to apply this to a new object class would require the manual collection of data points to be repeated [62].

Unsupervised approaches of alignment are described by common features of the face that can be identified by using different techniques such as Active Appearance Models (AAMs) [63], Haar feature detectors [58] and congealing methods [62]. Pfister et al. [64] used a combination of an Active Shape Model (ASM) [65] and Haar eye detection techniques as alignment methods to further improve normalisation. The final step of normalisation involves mapping the feature points of the face to the model face and outputting a result that is consistent with the position of the model face.

Shan et al. [66] outlines the problems with mis-alignment. Faces being misaligned can cause recognition problems due to faces being of different shapes, sizes and orientations. Other mis-alignment issues can be down to the user assuming that the input images and features being used have been accurately localised.

By using a deformable Lucas-Kanade algorithm, Zhu et al. [67] fit a mesh to faces that is automatically generated from a frontal face image based on the detected eye corners and the distance between the eyes. This template is then mapped onto the reference frame and the mesh is aligned to the input image with the help of the previous eye corner detectors. The face image is then rectified by mapping it onto the reference frame and can then be compared to the initial image and rectified image for differences.

This method has the limitation of producing some distorted faces due to the input face being at an odd angle or turned so far that when a mesh is fitted, the face becomes stretched and skewed. If looking for subtle changes in the face then this method could potentially make analysis of images inaccurate due to distortion to track the movements accurately. For normal facial detection, this method could still work by learning the overall feature map of a typical face.

2.6.3 Facial Data Extraction and Representation

There are two main areas of feature extraction used for facial expression analysis: feature-based and appearance-based. Geometric feature-based methods extract facial components or facial feature points and form a feature vector to represent the geometry of the face. These features map the shape and location of things such as the nose, eyebrows and mouth.

Appearance-based methods are designed to extract the appearance (skin texture) changes of the face, such as wrinkles and furrows. These features can be extracted on either the whole-face or specific regions in a face image.

2.6.3.1 Feature-Based Methods

By modelling the skin and muscle structure, Essa and Pentland [1] are able to use optical flow to estimate the muscle control variables that then classify facial expressions. This method also allows for an avatar-like representation of the face showing individual muscle group activation. By modelling temporal changes, this early work on facial expression recognition and representation did not require machine learning, but was able to plot several muscle movements into groups describing the onset, apex and offset of the expression.

Not all methods attempt to recognise expressions, AUs or emotion classes. Cohn et al. [21] proposed an AAM [63] based facial action measurement system to detect depression compared with clinical diagnosis. Manual FACS annotation and pitch analysis were also incorporated into the study. The automatic AAM system and manual FACS used leave-one-out validation with an SVM classifier and achieved 88% and 79% accuracy respectively. Logistic regression was used for determining vocal patterns of participants and achieved an accuracy of 79%. As this method shows, facial expression recognition has valid clinical applications, even against the normal tests performed.

Li et al. [68] proposed the use a hierarchical framework based on a Dynamic Bayesian Network to recognise facial expressions. Using facial point tracking, expressions are able to be modelled to best represent a particular facial movement. This method is able to recognise AUs and six of the basic expressions (excluding contempt).

2.6.3.2 Appearance-Based Methods

Littlewort et al. [69] proposed a fully automatic system for the real-time recognition of basic emotional expressions (neutral, anger, disgust, fear, joy, sadness, surprise) from posed expression video sequences [30]. The system codes each frame into one of the seven emotions (and neutral) by preprocessing using a bank of Gabor filters. A two-stage classification process is then used to solve the problem of unbalanced datasets. Stage one used SVM and trained on every possible pair of emotions. The goal of the second stage is to convert the first stage representation into a probability distribution over 7 expression categories. Multinomial logistic ridge regression (MLR) was used to complete this stage.

Instead of using posed facial expressions, Bartlett et al. [70] employed a user independent fully automatic system for real time recognition of FACS AUs from the RU-FACS dataset [71]. Each frame is coded in respect to 20 AUs with results obtained by selecting a subset of Gabor filters [72] using AdaBoost and then training SVM on the outputs of the filters selected by AdaBoost. Detecting and recognising spontaneous facial expressions is more useful for real-time application, as they represent genuine feelings of emotion.

Fan and Tjahjadi [73] expand pyramid histogram of gradients (PHOG) [74] into three orthogonal planes (PHOG-TOP), which leads to a facial expression recognition framework that uses dynamic information. The fusion of dense optical flow and PHOG-TOP is then implemented to detect facial movement in four sub-regions (forehead, mouth, eyebrow and nose). The datasets used was the extended Cohn-Kadade dataset [31] and MMI dataset [32].

2.6.3.3 Hybrid Methods

Tian et al. [75] use facial feature tracking and convert the results to detailed parametric descriptions of the facial features. With these features as the inputs, 11 lower face AUs and 7 upper face AUs are recognized by a neural network (NN) algorithm. The proposed method takes an objective approach to recognising facial expressions by not classifying into the prototypic emotions, but using FACS-based muscle movements to model the wrinkles and furrows that occur when a person exhibits facial expressions.

Facial surfaces usually contain person-dependent details such as facial hair, permanent wrinkles, scars and freckles. Appearance features alone can be dependent on these details for classification. Wen et al. [76] capture subtle facial motions in 3-Dimensional (3D) non-rigid face tracking to remove this dependence and can adapt recognition models to fit new people using individualised exemplars. The adaptive method is completed using an online Expectation-Maximization (EM)-based algorithm and classified using a Gaussian Mixture Model. Seven exemplars are used and represent six of the basic emotions (anger, disgust, fear, happiness, sadness, and surprise) and neutral using videos from posed expression data [30].

2.6.4 Facial Expression Recognition

The recognition problem for facial expressions is usually split into two categories: frame-based and sequence-based. To summarise, the first does not use temporal information and the latter does. A summary of techniques and results can be seen in Table 2.3, where recent recognition methods are outlined using different publicly available datasets. Although not exhaustive, this list was chosen to show a variety of approaches for feature extraction, recognition methods and use of facial expression datasets.

TABLE 2.3: A summary of some facial expression recognition methods with the corresponding feature, accuracy and dataset used.

System	Feature	Recognition Method	Recognition Rate	Databases
[77]	Gabor-wavelets	Neutral network	95.5%	Ekman-Hager/ Cohn-Kanade [30, 78]
[79]	LBQ-TOP	GentleBoost+ HMM+SVM	86.5%	MMI [32]
[80]	Bag of Words	SVM	96%	CK+ [31]
[81]	Gabor-wavelets	Similarity-based	67.9%	Jaffe [81]
[82]	AAM	SVM	78.6%	SEMAINE [83]
[84]	PCA-LDA	SVM	82.1%	Belfast Induced [85]
[86]	Gabor-filters	Robust Metric Learning	94.4%	MFP [87]

For macro-facial expressions, the usual method of classification is machine learning. Methods that have been applied include neural networks (NN), SVM [12], linear discriminant analysis (LDA), K-nearest neighbor [88], multinomial logistic ridge regression (MLR) and hidden Markov models (HMM) [89].

Frame-based methods use the information from the current input frame with or without a reference frame. The reference frame is commonly a neutral face to recognise the facial expression has occurred. Some frame-based methods include [69, 75–77, 90, 91]. By not using temporal information to inform about the person, the systems do not rely on an individual's neutral face.

By using temporal information, sequence-based approaches aim to recognise facial expressions in one or more frames. Some examples of sequence-based methods can be found in [3, 70, 92], with some being able to tackle the problem of recognising spontaneous expressions instead of posed.

2.6.5 Micro-Facial Expression Analysis

Unlike normal facial expressions, micro-facial expression analysis is less established, but is growing in popularity in recent years. This area has even been popularised to the general public with television series 'Lie to Me' and feature

film 'Inside Out'. Being able to detect these expressions with similar accuracy to state-of-the-art facial expression recognition methods is the one of the biggest challenges facing this field.

Micro-facial movements are difficult to spot for humans, so computer vision is used to try and detect the subtle changes we might miss. Unfortunately, unlike the large movements of macro-expression, micro-movements can be so small that computer vision systems can find it difficult to differentiate a true expression and noise (head movement, lighting changes).

The pipeline of detecting or recognising micro-movements is very similar to facial expression analysis in terms of acquisition and, if applicable, machine learning classification. However, geometric feature-based methods are rarely used as tracking feature points on a face that barely moves will not produce good results. Instead, appearance-based features are primarily used to attempt to describe the micro-movement or train machine learning to classify micro-expressions into classes.

2.6.5.1 Face Regions

Recent work on the recognition of micro-facial expressions have provided promising results on successful detection techniques, however there is room for improvement. To begin detection, current approaches follow methods of extracting local feature information of the face by splitting the face into regions.

Shreve et al. [93] split the face into 4 quadrants and analyse each quarter as individual temporal sequences. The advantage of this method is that it is simple to analyse larger regions, however the information to retrieve from the areas are restricted to whether there was some form of movement in a more global area.

Another method is to split the face into a specific number of blocks [17, 94]. The movement on the face is analysed locally, rather than a global representation of the whole face, and can focus on small changes in very specific temporal blocks. A disadvantage to this method is that it is computationally expensive to process the whole images as $m \times n$ blocks. It can also include features around the edge of the face, including hair, that do not relate to movement but could still effect the final feature vector.

Delaunay triangulation has been used to form regions on just the face and can exclude hair and neck [95], however this approach can still extract areas of the face that would not be useful as a feature and adds further computational expense.

A more recent and less researched approach is to use defined Regions of Interest (ROI) to correspond with one or more FACS AUs [96, 97]. These regions have more focus on local parts of the face that move due to muscle activation. Unfortunately, currently defined regions do not cover all AUs and miss some potentially important movements such as AU5 (Upper Lid Raiser), AU23 (Lip Tightener) and AU31 (Jaw Clencher).

2.6.5.2 State of the Art Micro-Expression Recognition Methods

Pfister et al. [64] used a Temporal Interpolation Model (TIM) based on a Laplacian matrix to normalise frame numbers in spontaneous micro-expression clips. Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [94] is then used to extract the temporal features. This method produces decent results and compares the existing York Deception Detection Test (YorkDDT) dataset to their own. This comparison is not directly comparable due to the datasets being very different, for example, the frame rate varied and resolution was low compared to the 640×480 of the Spontaneous Micro-Expression Corpus (SMIC) dataset. Statistically, the TIM model is able to make results of very short expression stable, however using interpolation, by adding frames that did not exist, could skew the already sensitive micro-expression data.

HOG features [11] was originally created for human detection in 2-Dimensional (2D) images and used the pixel orientation values, weighted by its magnitude, to calculate features for describing a human as an object. Polikovsky et al. [5, 98] used a 3D gradient histogram descriptor to recognise micro-facial expressions from high-speed videos. The paper used manually marked up areas that are relevant to FACS [39] based movement so that unnecessary parts of the face are left out. This does means that the method of classifying movement in these subjectively selected areas is time-consuming and would not suit a real-time application like interrogation. The spatio-temporal domain is explored highlighting the importance of the temporal plane in micro-expressions, however the bin selection for the XY plane is 8 and the XT, YT planes have been set to 12. The 8 bins in the XY planes was said to represent the different directions of movement, however the authors

do not justify the temporal bin selection. An improvement on this would be to gradually increase bin size to find the ideal size for this purpose using 3D gradient descriptors.

Wu et al. [99] use Gabor filters to automatically detect when recognising a micro-expression and spotting when they occur. However the training data used was 48 videos from Micro-Expression Training Tool (METT) [100] which flashes an expression very quickly to the user, but is classed as a posed expression. These changes may be quick but frame by frame analysis should be able to pick up these unnatural movements easily. The use of non-spontaneous micro-expression can explain how the recognition accuracy reached 95.8%.

Micro-expressions may be able to be used to help detect deception [41, 101], so Owayjan et al. [102] developed a system designed to directly detect lies with a reported accuracy of 85%. The system is designed and implemented using labVIEW and an embedded vision system that captures participant's interviews. By comparing with trained templates of specific expressions, the system finds the distances between features found on the participant's face to determine if they are lying after being asked some interview questions. The system makes large conclusions to what they deem to be a lie detection system, as even 100% detected micro-expression in isolation would not be truly accurate. Also, the system was only tested on 4 participants and as the system tried to find subtle distance changes in facial expressions, participants would be answering questions and therefore making their faces move considerably.

Song et al. [103] attempt to learn a 'codebook' of micro-expressions from local space-time features and obtain a sparse representation of these features. After the codebook is obtained, a prediction model, namely support vector regression, is used to infer an emotional state. The features used are spatio-temporal interest points (STIP) that find local feature patches in a temporal sequence using HOG/HOOF features. The dataset used in experiments was not a micro-expression dataset, but contained emotional sequences that were used to obtain the face features that were deemed to be micro-expressions. As the method does not go into detail about the data, it is assumed the movements were not FACS coded.

A simple method is proposed by Guo et al. [104] uses LBP-TOP and the nearest neighbour algorithm for classification. Using the SMIC [15] dataset, the method extracts features using the whole face, and does not split the images into

smaller blocks or regions. The amount of sequences used was 70 negative and 51 positive. After splitting the sequences into testing and training sets, the Euclidean distance is calculated between various ratios of testing and training data to get a recognition accuracy. with a ratio of 5:1 (training:testing) the method achieved 63% accuracy. The accuracy score reflects similarly for other methods, and shows the difficult nature of recognising micro-expressions. Also, by not splitting the face into more local regions, micro-expressions are harder to distinguish.

Wang et al. [96, 105] used grey-scale videos clips in the 3rd-order tensor and used discriminant tensor subspace analysis. Further, they use a tensor independent color space (TICS) model to show performance of micro-expression recognition in a different colour space compared with RGB and grey-scale. Both papers normalises the face to a model with an active shape model and is effective in making the shape and location of faces consistent, however deforming faces to a model face could interfere with micro-expression recognition and potentially remove subtle movements. The recognition methodology used a cut-down version of Pfister et al. [64] by only using LBP-TOP features to a selective image block size of 5×5 and 8×8. No justification was given to why these block sizes were selected. In addition, the second paper (using TICS) contributes a novel colour representation of micro-expressions rather than grey-scale or RGB values. Unfortunately, the highest results from this colour space representation are 58.64%, which is only a few percent better than the RGB (56.09%) and grey-scale (56.91%).

A newly proposed feature, Main Directional Mean Optical-flow (MDMO), has been developed by Liu et al. [8] for micro-facial expression recognition using SVM as a classifier. The method of detection also uses 36 regions, partitioned using 66 facial points on the face, to isolate local areas for analysis, but keeping the feature vector small for computational efficiency. The best result on the CASME II dataset was 67.37% using leave-one-subject-out cross validation, which performed better than the LBP-TOP and Histogram of Oriented Optical Flow (HOOF) features. The results are similar to previous methods, but none yet have achieved over 90% accuracy seen in macro-facial expression analysis methods. Further, as the MDMO feature uses SVM for classification, the vector must be normalised and therefore loses the frame-based temporal attributes that would be useful for detecting onset, apex and offset frames.

Wang et al. [106] proposed two feature extraction approaches, and then used SVM with a linear and radial basis function (RBF) kernels for classification. The

first feature was an expansion to the LBP-TOP spatio-temporal feature, where they removed computed neighbour points from the three orthogonal planes by considering only six unique intersection points on the three intersecting lines of the three orthogonal planes. The second feature computed the mean image of each video volume stack of orthogonal planes, therefore only deriving three mean images for each video correspond to the XY, XT and YT planes. The features were compared on the SMIC [15] and CASME II [17] datasets. One of the highest results was 66.40% on a 'leave-one-video-out' approach, using an RBF kernel, using the six intersection points and tested on the CASME II dataset. Unfortunately, the results do not go higher than around 70%, and the six intersecting points feature took around 15 seconds to extract each feature.

Lu et al. [95] proposed a Delaunay-based temporal coding model, which recognises micro-expressions by encoding texture variations and relating them to muscle movements. By using Delaunay triangulation, participants in a dataset can be normalised to reduce the influence of personal appearance. The use of triangulation allows for regions to be defined and used to calculate the magnitudes within the regions to select features that are related to micro-expressions. The experiments used the SVM and RF algorithms for classification, but the best result was achieved when trying to separate micro-expressions and non-movements with an accuracy of 88.89%. By using local regions, subtle movements in small areas can be localised. However, by using Delaunay triangulation across the whole face the features may still include irrelevant information in the final feature vector, similar to block-based approaches [16, 17].

2.7 Micro-Movement Detection

Research into detecting micro-movements rather than the recognition of micro-expressions using machine learning is very limited. The focus on recognition approaches, as discussed in Section 2.6.5.2, is a result of assuming the task of putting micro-expressions into categories is trivial. The results from many of these studies show very few achieving higher than chance, including studies way below 50% accuracy. Going against the machine learning approach can be controversial, but objectifying micro-movements as skin deformations of the face could lead to better detection methods and finally a better understanding of micro-facial movements.

Shreve et al. [6, 93, 107] proposed a novel solution of segmenting macro- and micro-expression frames in video sequences by calculating the strain magnitude in an optical flow field corresponding to the elastic deformation of facial skin tissue. The author's techniques of using strain magnitude is the most natural way of calculating whether a micro-expression has occurred as this is how a human would interpret using their visual system. The thresholding technique used to determine a macro-expression is also used for micro-expressions, however it is more constrained due to their rapid speed and spatial locality on the face. First, macro-expressions are removed using the FE detection algorithm. Next, two additional criteria are added:

1. the strain magnitude has to be larger than surrounding regions
2. the duration of this increased strain can be no more than 1/5th second.

As with others in this field, the limitation of datasets means that this paper could not complete a comprehensive evaluation of available data, and used the University of South Florida-High Definition (USF-HD) [6], Canal-9 [108] datasets and found videos from the Internet. None of these had a high-speed frame rate. The ground truth coding on the datasets have not been performed by trained FACS coders, therefore the reliability and consistency of knowing what is and is not an expression cannot be certain. The paper does not use any temporal methods due to the use of optical flow and spatial skin deformation, results in a 44% false positive rate for detecting micro-expressions.

Moilanen et al. [109] used an appearance-based feature difference analysis method that incorporates chi-squared (χ^2) distance and peak detection to determine when a movement crosses a threshold and can be classed as a movement. This follows a more objective method that does not require machine learning. The datasets used are the CASME-A and B [16] and the original data from the SMIC [15] (not currently publicly accessible). For CASME-A the spotting accuracy (True Positive Rate (TPR)) was 52% with 30 False Positives (FPs), CASME-B had 66% with 32 FPs and SMIC-VIS-Extended(E) achieved 71% with 23 FPs. The threshold value for peak detection was set semi-automatically, with a percentage value between [0,1] being manually set for each dataset. Only spatial appearance is used for descriptor calculation, therefore leaving out temporal planes associated with video volumes.

By exploiting the feature difference contrast, Li et al. [9] proposed an algorithm that spots micro-expressions in videos. Further, they combine this with an automatic micro-expression analysis system to recognise expressions of the SMIC-VIS-E dataset in one of three categories: positive, negative or surprise. This spotting algorithm is very similar to the feature difference method seen in [109], however HOOF is compared along with Local Binary Patterns (LBP). The highest result came from using LBP in the CASME II dataset with an AUC of 92.98%. The best performance came from using the SMIC-VIS-E dataset and LBP, where the spotting accuracy was about 70% of micro-expressions, 13.5% False Positive Rate (FPR), and the AUC was 84.53%.

The proposed micro-expression spotting system in [9] still uses a block-based approach, as in [109], therefore potentially including redundant information. There is also no face alignment performed for these video clips which could lead to head movements being falsely spotted. It should also be noted that throughout this paper, the use of 'accuracy' is not defined. Using more reliable statistical measures, such as recall, precision and F-measure, would allow the system to be scrutinised more effectively. Block-based approaches, proposed in [9, 109], split the face into $m \times n$ blocks. Doing this can include a lot of redundant information, in other words, non-muscle movement. In addition, there are no indication of where on the face the movement occurs. A need for a better solution is required to focus on areas of the face that provide important information while reducing computational complexity.

Further, [9] test untrained humans to spot micro-expressions from a selected 71 clips from the SMIC-VIS dataset. The result of the 15 participant's spotting accuracy was 71.11% (Standard Deviation (SD) = 7.22%), however for untrained participants this seems very high considering other studies have found many participants can only recognise around 50% on average [110, 111]. The second human study in [9], of mainly spotting when micro-expression occurred, appears more usual with a result of 49.74%.

Xia et al. [112] proposed a probabilistic framework with random walk algorithm to detect spontaneous micro-expression clips temporally from a video sequence. Geometric deformation is captured by an ASM model [65] and is utilised as features which are robust to subtle head movement and illumination variation. The Adaboost algorithm is then used to estimate the initial probability for each frame and compute the transition probability by deformation similarity. The

method is validated on the SMIC [15] and CASME dataset [16] where it achieved a spotting accuracy of 0.8693 and 0.9208 respectively. It should be noted that the newer CASME II [17] is not used despite being newer and contains more data.

Patel et al. [113] introduced a method using optical flow motion vectors, calculated within small ROI built around facial landmarks, to detect the onset, apex and offset of a micro-movement. The first step detects 49 facial points, with the tracking of points over subsequent frames being calculated using optical flow vectors. Small ROI are created around facial landmarks, creating 8 ROI grouped around certain points. These are use to balance the inaccuracy of landmark detection and the groups correspond to AU in FACS. Finally, the method can remove head movements, eye blinks and eye gaze changes, common reasons for false positives in micro-movement detection methods, by the use of thresholding. The motion for an AU activation should be high, but low for other points. A peak frame is considered true if all the points of an AU group have a motion greater than a certain threshold. An attempt is made to get this system to perform in real-time, however many of the computational time are in seconds, including the facial landmark detection and optical flow calculation. The method also only uses the SMIC dataset at 25 fps, which means the micro-movements are not FACS coded and has a limited temporal resolution for finding subtle motions. The computational times also take a long time at this frame rate, and so higher frame rates for this method would be even higher. The results detailed an AUC of 95%, but produced a high number of FPs.

2.8 Current Benchmark Datasets

2.8.1 Polikovsky Dataset

One of the first micro-expression datasets was created by Polikovsky et al. [5]. The participants were 10 university students in a laboratory setting and their faces were recorded at 200 fps with a resolution of 640×480. The demographic was reasonably spread but limited in number with 5 Asians, 4 Caucasians and 1 Indian participant.

The laboratory setting was set up to maximise the focus on the face, and followed the recommendations of [114]. To reduce shadowing, lights were placed

above, to the left and right of the participant. The background consisted of a uniform colour of approximately 18% grey. The camera was also rotated 90 degrees to increase the pixels available for face acquisition.

The micro-expressions in this dataset were posed by participants whom were asked to perform the 7 basic emotions. Posed facial expressions have been found to have significant differences to spontaneous expressions [115], therefore the micro-expressions in this dataset are not representative of natural human behaviour and highlights the requirement for expressions induced naturally. Further, this dataset is not publicly available for further study.

2.8.2 USF-HD

A similar dataset was created named USF-HD [6] and includes 100 posed micro-expressions recorded at 29.7 fps. The participants were shown various micro-facial expressions and told to replicate them in any order they wished. As with the previously described dataset, posed micro-expressions do not re-create a real-world scenario and replicating other people's micro-expressions does not represent how these movements would be presented by the participants themselves.

Recording at almost 30 fps can risk losing important information about the movements. In addition, this dataset defined micro-expressions as no higher than 660 ms, which is longer than the previously accepted definitions. Moreover, the categories for micro-expressions are smile, surprise, anger and sad, which is reduced from the 7 universal expressions by missing out disgust, fear and contempt. This dataset has also not been made available for public research use.

2.8.3 YorkDDT

As part of a psychological study, the York Deception Detection Test (YorkDDT), Warren et al. [116] recorded 20 video clips, at 320×240 resolution and 25 fps, where participants truthfully or deceptively described two film clips that were either classed as emotional, or non-emotional. The emotional clip, intended to be stressful, was of an unpleasant surgical operation. The non-emotional clip was meant to be neutral, showing a pleasant landscape scene.

FIGURE 2.4: An example of a 'positive' labelled micro-movement from the SMIC dataset.

The participants watching the emotional clip were asked to describe the non-emotional video, and the opposite for the participants watching the non-emotional clip. Warren et al. [116] reported that some micro-facial expressions occurred during both scenarios, however these movements were not reported to be available for public use.

During their study into micro-facial expression recognition, Pfister et al. [64] managed to obtain the original scenario videos where 9 participants (3 male and 6 female) displayed micro-expressions. They extracted 18 micro-facial expressions for analysis, 7 from the emotional scenario and 11 from the non-emotional version.

Other than the very low amount of micro-expressions in this dataset, it is created through a second source that do not go into a large amount of detail about AUs, or participant demographic. With the data unable to be publicly accessed, it is not possible to study these micro-expressions. It is also an issue with the frame rate being so low, the largest amount of frames for analysis would be around 12-13 frames. The lowest reported micro-expression length was 7 frames.

2.8.4 SMIC

The SMIC dataset [15] consists of 77 spontaneous micro-facial expressions filmed at 100 fps and was one of the first to include spontaneous micro-expressions obtained through emotional inducement experiments. However, this dataset was not coded using FACS [39] and gives no information on neutral sequences (the participant's face not moving before onset).

The protocols for the inducement experiment consisted of showing participants videos to react to and asking them to suppress their emotions, however with no FACS coding the categorisation of emotion labels was left to participant's own

self-reporting by asking to fill out how they felt during each video. Leaving the categorisation to participants allows for subjectivity on the emotional stimuli to be introduced. The recording quality was also decreased due to flickering of light in the laboratory setting and the facial area was 190×230 pixels.

The SMIC included a wider demographic of participants with 6 being female and 14 male. Ethnicity was more diverse than previous datasets with ten Asians, nine Caucasians and one African participant, however this still only includes 3 ethnicities and does not provide a good overview of a population. This dataset also includes recordings at 25 fps, and near-infrared images but for consistency the high-speed videos are the only ones used for comparison. An example from the dataset can be seen in Fig. 2.4 that shows a 'positive' micro-movement.

2.8.5 CASME

To address the low number of micro-expressions in previous datasets, the CASME dataset [16] captured 195 spontaneous micro-expressions at 60 fps, but the facial area dimensions were lower than SMIC at 150×190 pixels. A further advantage of this dataset is all expressions are FACS coded and included the onset, apex and offset frame number. The duration of any movement did not exceed 500 ms unless the onset duration was less than 250 ms because fast-onset facial expressions, with slow offset durations, are classed as micro-expressions [43].

Participants watched videos to induce an emotional response. All 17 videos were selected from the with a positive and negative valence. The authors acknowledge that the videos may elicit various emotions from the videos. Video durations ranged from 1 - 4 minutes, and mainly were used to elicit 1 particular emotion. Finally, participants rated the videos afterwards on the emotional content from a scale of 0 - 6, where 0 was the weakest and 6 the strongest.

The frame rate of this dataset, 60 fps, does not represent micro-expressions well, as the movements could easily be missed when recording. The categories for classifying a labelled emotion have been selected based on the video content, self-report of participants and universal emotion theory. Moreover, the dataset uses repression and tense as new additions aside from the universal emotion theory and leaves out contempt and anger.

FIGURE 2.5: An example micro-movement from the CASME II dataset. In static images, it is not easy to see the movement, however the participant has been FACS coded with AU12 (lip corner puller).

2.8.6 CASME II

Shortly after, CASME II [17] was created as an extension of the original CASME dataset. The frame rate increased to 200 fps to analyse more detail in muscle movements, and 247 newly FACS coded micro-expressions from 26 participants were obtained. The facial area used for analysis was the larger than CASME and SMIC at 280×340 pixels. An example from the dataset can be seen in Fig. 2.5 that shows AU12 (lip corner puller) induced from this participant.

However, as with the previous version, this dataset includes only Chinese participants and categorises in the same way. Both CASME and CASME II used 35 participants, mostly students with a mean age of 22.03 (SD = 1.60). Along with only using one ethnicity, both these datasets have the disadvantage of using young participants only, restricting the dataset to analysing similar looking participants (based on age features).

Based on the findings from the previous benchmark datasets, much more can be done to address the limitations such as consistent lighting and a wide demographic, however the lack of datasets for micro-movements induced spontaneously motivates the creation of this dataset.

2.9 Real-World Applications

Detecting and recognising human faces and facial expressions has attracted increasing attention from psychologists, cognitive scientists and computer scientists due to its application to human computer interaction [117, 118], law enforcement and security [119] and computer animation.

Successfully detecting micro-facial movements to a high degree of accuracy has benefits to many areas, including security, psychological assessment, teaching and training [13, 19, 20]. Within airports the use of such technology can spot a face in the crowd who may be concealing emotion and showing micro-movements. Spotting this early can help prevent a person potentially causing harm as they try to conceal the intent in their facial expressions. This system may be a long way off yet due to the technological challenges of spotting such small movement with security cameras in an area with a dense population.

It should be emphasised that systems in security would not be able to immediately spot if someone wants to cause harm, but will provide a starting point to investigate further. This links with another application of automatic micro-movement detection where people can be taught how to spot the expressions much easier if they can review situations where they occur, such as interrogations of criminals or experiments to spot liars.

A lot of applications show promise in security and law enforcement, where most research focuses applications, however there is a possibility to apply micro-facial movement detection to psychological assessments and neurological problems. A person having problems with their mind may try to conceal their true feelings to others, but they are truly under distress from problems arising from their lives [21].

Mental illnesses show few physical symptoms, but patients yet to be diagnosed with a mental health issue could be helped faster if their concealed distress is uncovered. Further, neurological issues in the brain may cause involuntary twitches on the face that could show early signs of diseases like Parkinson's [22]. Therefore, even though the twitches will not be related the facial expressions it shows the broader scope of the solution as an application to real-world problems.

Paul Ekman's research was used by the United States' Transportation Security Administration (TSA) to train officers on how to analyse behaviour for border protection [14]. It is also known that this research has been used to train large organisation for crime prevention such as the FBI and CIA, showing that this research is being recognised as important for future crime prevention.

2.10 Research Direction

Analysing large or macro-facial expressions has been a big part of computer vision for decades, and the work done here seemingly can be replicated for micro-facial expression analysis. However, this field has branched into an area of its own garnering a multitude of experiments and applications [10, 13, 18–20, 43, 46, 100].

With global threats such as terrorism and flawed science such as the polygraph for deception detection [120], using the non-verbal cue of micro-expressions may be a key to helping aid security in dealing with potential threats. Security and deception detection are not the only application though, as using systems to help in the detection of micro-expressions could help for clinical applications such as helping therapists uncover hidden feelings in patient's regardless of their awareness of the emotion. Other potential uses would be for early facial paresis detection [22], cross-culture studies [23, 24] and computer animation [25].

A controversial point is whether or not it should be allowed to detect these micro-expressions, as the theory behind it states that the person attempting to conceal their emotion experience these movements involuntarily and likely unknowingly. If we are able to detect them with high accuracy, then we are effectively robbing a person of being able to hide something that is private to them. From an ethical point of view, knowing when someone is being deceptive would be advantageous but takes away the freedom you had in your emotions.

Regardless of current issues surrounding privacy, micro-facial expression analysis systems are still in the early stages of development, with accuracies reaching roughly no more than 70% [5, 16, 17, 109]. Further, the methods have limited data to work with, especially compared to normal facial expression, due to the difficult nature of inducing micro-facial expressions. Also, as the micro-movements are so fast, normal camera systems that record between 25 - 30 fps barely capture any useful information for micro-movement analysis. High-speed cameras are therefore used [15, 17], which captures many more frames (100 - 200 fps) but at the expense of any current system not being able to perform in real-time.

For any micro-movement detection algorithm, there can be many problems to solve to ensure an effective detection system. Based on the review of current literature, a few research challenges are identified. The speed and subtlety of micro-movements makes differentiating between micro-movements and non-movements

challenging, especially ones that are genuine and spontaneous. Similarly, image noise can affect detection algorithms considerably. The use of machine learning has been explored, however recognising micro-movements into distinct classes becomes challenging when micro-movements can be expressed in many small facial muscle motions. Further, machine learning requires training with a large, balanced dataset and is not suitable for real-time processing.

2.11 Summary

This Chapter introduced the theory of universal facial expressions and why they are important in understanding and interpretation of emotion. It moves on to discuss facial expression analysis that informs the main review of micro-facial expression analysis. Analysing these subtle expressions is difficult for humans, but approaches to representing these movements using computer vision are steadily growing, but have a long way to go before being as well-established as facial expression analysis.

Micro-facial expression analysis tends to focus on the emotional interpretation, meaning assumptions are commonly made. Focusing on micro-facial movements, which describe the facial muscle activations, removes these assumptions. This Chapter described the way machine learning makes many assumptions in the recognition of micro-expression into distinct categories, and in contrast, micro-movement analysis and detection is described as a more suitable approach when using computer vision.

Chapter 3

Theories and Techniques

This Chapter focuses on the techniques used throughout the thesis, including detecting the face, alignment of face into a canonical plane, de-noising, feature extraction and measuring the performance of results. The flow of this Chapter loosely follows the flow in which many algorithms take during the analysis of micro-facial movements.

3.1 Face Alignment

3.1.1 Face Detection

The Viola-Jones algorithm is a form of machine learning for object detection that is able to process images rapidly to achieve high detection rates [58]. A popular use for this algorithm is for face detection. The overall approach details three main steps in object or face detection: selection of Haar-like features and calculating the integral image; feature selection by Adaboost [121]; and combination of weak classifiers to form the final strong classifier. Further explanation of this method can be found in Appendix A.

Haar features can also be used to train individual parts of the face such as the eyes, mouth or nose. By using the face detector and eye detector together, faces can be rigidly rotated using an affine transform (see Section 3.1.3) to normalise the face, or multiple faces, in respect to the positions of the eye coordinates. Fig. 3.1

FIGURE 3.1: The face is acquired continuously through a webcam, and the face and eyes are detected using Haar features. Using the detected features the face can be cropped and aligned as can be seen in the upper part of the image.

shows the process of detecting the face, the eyes and finally cropping and rotating the face in real-time using an inexpensive webcam.

3.1.2 Facial Landmark Detection

Faces have quite distinct features, such as eyebrows, mouths, noses and eyes. Most humans have these features in about the same position, and so identifying the points where these features occur on the face is an interest research problem.

Zhou et al. [122] proposed a way of detecting facial points using a deep learning approach named convolutional neural networks (CNN). Two advantages of using a CNN are the geometric constraints are implicitly utilised and a large amount of data can be used for training. The research toolkit developed requires an Internet connection to allow for the images to be processed on the Face++ servers. However, the uses range from predicting age and gender to facial expression estimation.

The system won the industry prize in the 300 Faces In-the-Wild Challenge (300-W) [123] achieving a median absolute deviation score of 0.0205 during fitting tests. The system used four faces-in-the-wild datasets to train the CNN : AFW, LFPW, HELEN, and iBUG [124–128]. The use of a combination of these datasets allows for a more robustly trained system.

The system design is a four-level convolutional network cascade, and handles the problem in a coarse-to-fine manner. The network takes the raw pixels as input (the face images) and performs regression on the coordinates of the desired points. The CNN is composed of multiple linear and non-linear operators.

$$f_{CNN}(x) = f_n(f_{n-1}(\dots f_1 \dots (x))) \tag{3.1}$$

The first operator is the convolution layer which filters the multi-channel image signal and is defined as

$$C_W(x)_{i,j,k} = \sum_{u,v,w} W_{u,v,k,w} x_{i-u,j-v,w} + B_k \tag{3.2}$$

The second operator performs max-pooling, which reduces the size of images by

$$M_s(x)_{i,j,k} = \max_{0 \le u,v < s} x_{is-u,js-v,k} \tag{3.3}$$

The third and final is a non-linear operator of unshared convolution layers and is defined as

$$g(x) = |\tanh(x)| \tag{3.4}$$

The three operator layers described, convolution, max-pooling and non-linear, are the three key modules that make up a CNN. An extension into multiple layers can be achieved by repeating the operators as needed. An example of the 83 facial points detected using Face++ can be seen in Fig. 3.2.

The main reason for using Face++ in this thesis is the robust nature of the system by using a large amount of training data [122, 123, 129]. However, many other facial landmark detection techniques exist. AAM was proposed by Cootes et al. [63] and can generate the shape and appearance of deformable objects. However, compared with Constrained Local Models (CLMs) [130, 131], an AAM is not as suited to real-time applications. Tzimiropoulos and Pantic [132] proposed Gauss-Newton Deformable Part Models, which combines Gauss-Newton

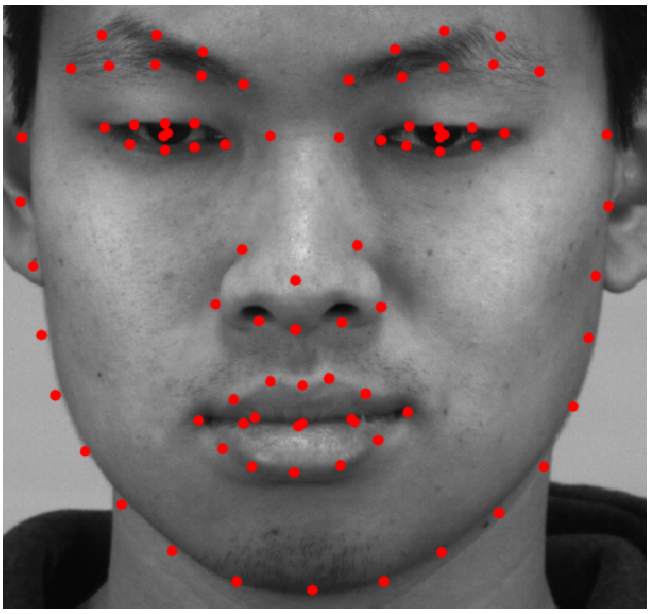

FIGURE 3.2: 83 points detected by Face++. It is also possible to reduce the amount of points calculated to 5 or 25, if necessary.

optimisation to minimise a joint cost function of the shape and appearance with Deformable Part Models (DPMs). The combination helps overcome the detection problems and limitations of DPMs on their own and achieves comparable results with [122]. Although these systems are suitable for this thesis, the following landmark detection methods have performed well in real-time. Zhang et al. [133] create an unsupervised automatic point detection method that is robust to rotations and occlusions and does not need large training sets like Face++. However, the average error achieved was 0.80, lower than Face++. Orozco et al. [134] created an on-line appearance-based tracker that could simultaneous track 3D head pose, lips, eyebrows, eyelids and irises in monocular video sequences. Similar to [133], this tracker does not require large training sets.

3.1.3 Affine Transformation

To transform a 2D geometric shape, while preserving lines, planes and points, affine transformation can be used. Any points (x, y) can be transformed into a new set of coordinates (x', y') by translating, rotating, scaling or shearing. A combination of two or more of these transformations can also be applied.

Each vertex of the object has the transformation applied instead of applying it to every point on every line that makes up the shape. New lines can then be

drawn between the resulting endpoints of the transformation. To move an object from one position to another, translation is used and is defined as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_x \\ 0 & 0 & 1 \end{bmatrix} \bullet \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.5}$$

where $\begin{bmatrix} x & y & 1 \end{bmatrix}^T$ are the homogeneous coordinates used to express the matrix multiplication in the 3×3 transformation matrix. The translation factor, t, is adjusted to change the location in x and y.

Rotation of an object causes the object to rotate around a point, usually the central point of the object or coordinate system origin, without changing the distance from that point. Rotating the object clockwise is defined as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \bullet \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.6}$$

and anti-clockwise as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \bullet \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.7}$$

where θ is the angle value for the object to be rotated. The range of rotation is between $0 \leq \theta \leq 2\pi$ in radians, which is the equivalent to $0 - 360°$.

To change the size of an object, scaling is used and introduced with a scaling factor s. The transformation is defined as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \bullet \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.8}$$

where the scaling factor s_x increases the size of the object in the x axis if the value is greater than 1, and decreases in size if less than 1. This is the same of

the scaling factor s_y in the y axis. If $s_x = s_y = 1$ then the size of the object is unchanged.

The final affine transformation is to shear the object in the x and y direction. This transformation displaces each point in a fixed direction by an amount proportional to its signed distance from a line that is parallel to that direction. For example, a change in the x direction for a rectangle object will form a parallelogram. The transformation for shearing is defined as

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & sh_y & 0 \\ sh_x & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \bullet \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.9}$$

where sh_x and sh_y specify the shear factor along the x and y axis respectively.

3.1.4 Piecewise Affine Warping

In Section 3.1.3, we obtained landmark points for each face. To complete the warping of textures from control points to new points, Piecewise Affine (PWA) is used. This function does not require training as it does not fit points to the face like traditional AAMs. Although this thesis does not have an exact value for computational efficiency, Cootes and Kittipanya-ngam [135] define the basic computation speed of a basic AAM to be 172 ms, meaning that because PWA is only warping triangulated areas from original points to new points, it would not need to do as much processing, therefore making it faster.

If we require an image $\mathbf{I}$ to be warped to a new image $\mathbf{I}'$, n control points, x_i, are mapped to new points x_i'. To understand the process of image warping, Cootes and Taylor [136] describe the continuous vector valued mapping function, $\mathbf{f}$, to project each pixel of image $\mathbf{I}$ to the new image $\mathbf{I}'$. The mapping function $\mathbf{f}$ is defined as

$$\mathbf{f}(x_i) = x_i' \; \forall i = 1, \ldots, n \tag{3.10}$$

To avoid holes and interpolation issues, it is better to calculate the reverse mapping, $\mathbf{f}'$ taking the points of x_i' into x_i. For each pixel in the newly warped image, $\mathbf{I}'$, it can be determined where in the original image, $\mathbf{I}$, it came from and fill it in.

The mapping function $\mathbf{f}$ can also be broken down into a sum,

$$\mathbf{f}(x) = \sum_{i=1}^{n} f_i(x) x'_i \tag{3.11}$$

where the n continuous scalar valued functions f_i each satisfy

$$f_i(x_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

ensuring that $\mathbf{f}(x_i) = x'_i$.

One of the simplest forms of warping function is PWA [136], where each f_i is assumed to be linear in a local region and zero everywhere else. As an example, in the one dimensional case (where each $\mathbf{x}$ is a point on a line), suppose the control points are arranged in ascending order ($x_i < x_{i+1}$). To arrange $\mathbf{f}$ so that it will map a point $\mathbf{x}$, that is halfway between x_i and x_{i+1}, to a point halfway between x'_i and x'_{i+1}, the following setting is applied

$$f_i(x) = \begin{cases} (x - x_i)/(x_{i+1} - x_i), & \text{if } x \in [x_i, x_{i+1}] \text{ and } i < n \\ (x - x_i)/(x_i - x_{i-1}), & \text{if } x \in [x_{i-1}, x_i] \text{ and } i > 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.13}$$

where the control points in the region is between x_1 and x_n and the image can only be warped between these points.

The images are being warped in 2D, so triangulation is completed using the Delaunay method to partition the convex hull of the control points into a set of triangles. For the points within each triangle, an affine transformation is applied which uniquely maps the corners of the triangle to their new positions in $\mathbf{I}'$ (see Fig. 3.3 for an example of a single frame being warped). To illustrate is transformation, suppose $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ are the three vertices of a triangle within the convex hull. Any internal point of the triangle can be written as

$$\begin{aligned} \mathbf{x} &= \mathbf{x}_1 + \beta(\mathbf{x}_2 - \mathbf{x}_1) + \gamma(\mathbf{x}_3 - \mathbf{x}_1) \\ &= \alpha\mathbf{x}_1 + \beta\mathbf{x}_2 + \gamma\mathbf{x}_3 \end{aligned} \tag{3.14}$$

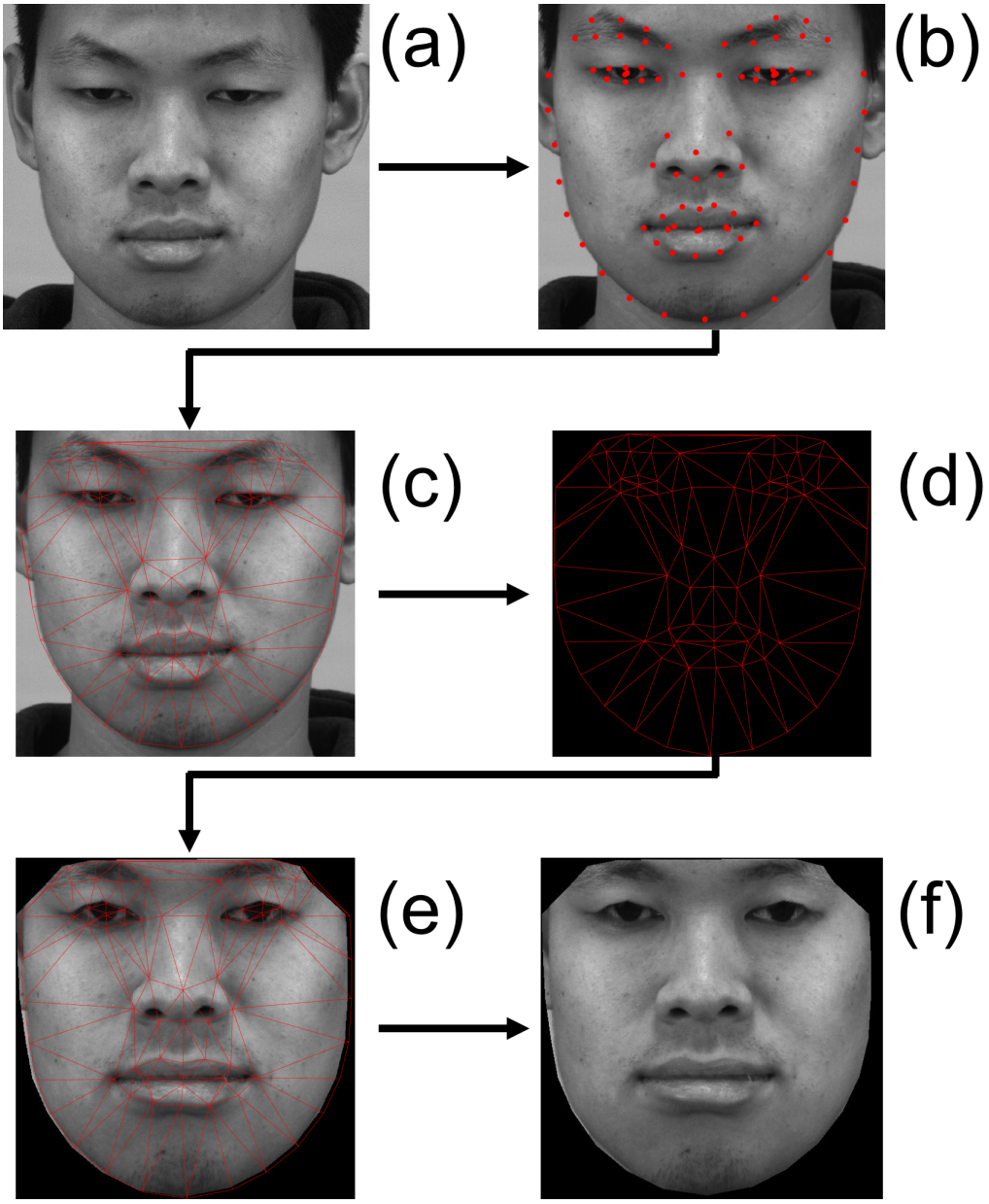

FIGURE 3.3: The process of piecewise affine warping. (a) the original image (b) the original face has points automatically detected (c) Delaunay triangulation creates the convex hull for to allow the mask to be warped (d) A template shape is defined that will allow the face to be warped to a particular shape (e) the warped face to the chosen shape (f) the final output. This process can be repeated for many faces to create a similar shape while keeping the participant's facial texture.

where $\alpha = 1 - (\beta + \gamma)$ and so $\alpha + \beta + \gamma = 1$. For $\mathbf{x}$ to be inside the triangle, $0 \leq \alpha, \beta, \gamma \leq 1$. Under the affine transformation, this point maps to

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}) = \alpha\mathbf{x}'_1 + \beta\mathbf{x}'_2 + \gamma\mathbf{x}'_3 \tag{3.15}$$

To create the warped image, each pixel $\mathbf{x}'$ in $\mathbf{I}'$ is set to the triangle it belongs to, the α, β and γ coefficients are computed to give the pixel's relative position in the triangle, which are then used to find the equivalent point in the original image, $\mathbf{I}$. The point is sampled and the value copied into pixel $\mathbf{x}'$ in $\mathbf{I}'$.

Although the PWA transformation gives continuous deformation, it is not smooth. This leads to straight lines being kinked across the triangle boundaries. Fortunately, face textures have few straight lines and are minimally effected by this issue. As every person has a different face shape, PWA transformation can generalise this shape while keeping the person's individual features and allows for consistent region analysis. Further, each warp for the defined triangles is performed once per triangle, not once per pixel. To ensure computational efficiency, the PWA transformation in this thesis follows the advice of Simon and Baker [137].

Traditional AAMs [63] would also be able to map texture from control points to new points, however as the points had already been located it is not necessary to use a method that would take longer to process and require training of a face model. As this thesis focuses on frontal view face images, PWA is very suitable, however if faces included rotations or profile views, it would not perform as well.

3.1.5 Subpixel Image Alignment via Fast Fourier Transform

To reduce the amount of rigid head movement that is inevitable when recording humans for long periods of time, each video frame of the micro-movement videos are aligned to be registered with the first frame of the sequence. As the micro-movements are subtle and last no more than 100 frames, the first frame is used as a reference frame.

To complete the alignment on our images without the need for facial points we use the method proposed by Guizar-Sicairos et al. [138]. This method uses a 2D-Discrete Fourier Transform (DFT) to calculate the cross-correlation and find

its peak, which can then be used to find a translation between the reference image, $f(x, y)$ and image to register, $g(x, y)$. It should be noted that all equations used in this thesis have been proposed by [138] and not by this research.

The first step is to apply the 2D-DFT to both images. To obtain subpixel accuracy, the resolution of $f(x, y)$ is upsampled from the image dimensions of M by N to $k * M$ by $k * N$, where k is an upsampling factor that defines a subpixel error of $1/k$.

An initial estimate of the translation to be applied to $g(x, y)$ is defined by $T(x, y)$ and is calculated by finding the cross-correlation peak using the inverse 2D-DFT and setting k to be 2. The 2D-DFT of the images are embedded into an array that is twice the size of the original image. The images can then be aligned to the estimated $T(x, y)$. This process continues until an upsampling factor of k is achieved.

By using a 2D-DFT, this method is able to select a local neighbourhood, around the initial peak estimate, to find the final peak calculated from cross-correlation rather wasting computing resources by calculating over the entire upsampled array. When using such large amount of images from high-speed video, processing efficiency is highly valued and provides a further step towards real-time micro-movement analysis.

To assess the accuracy of this method, Guizar-Sicairos et al. [138] corrupted a 256×256 complex valued image with additive zero-mean circular complex Gaussian noise and translated away from the original position. The error in estimated image shift Δr versus an upsampling factor k for $E = 0.25$ was calculated as $\Delta r = 0.0029$ pixels. The computation time was also calculated on a desktop computer with the specification: AMD Athlon X2 dual core processor 2.21 GHz, 64-bit operating system, 4 GB RAM. For an image size of 463×463 with k=25, the traditional FFT upsampling approach takes 235 s, as compared to 0.78 s with the DFT approach.

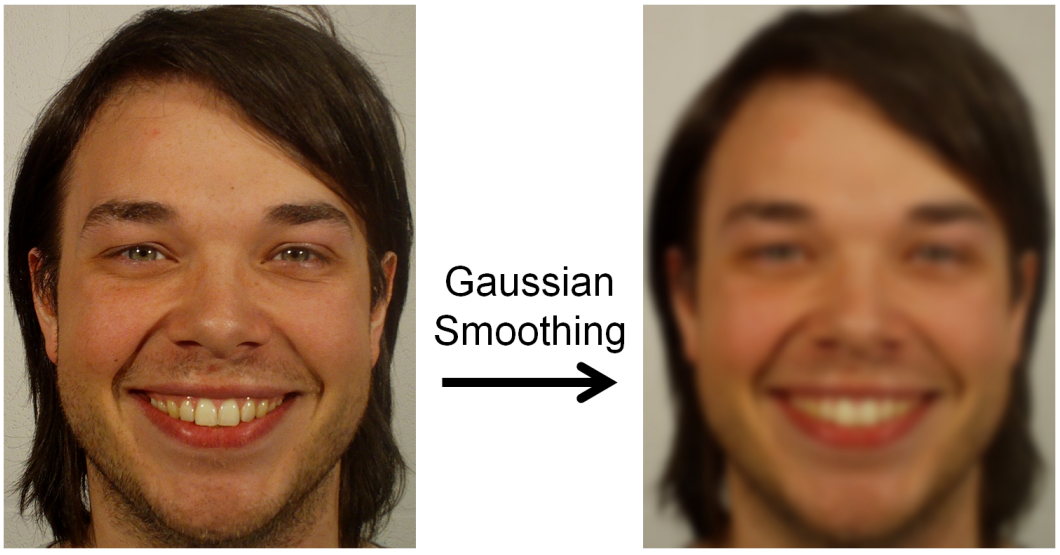

FIGURE 3.4: The original image on the left has the Gaussian smoothing operator applied. This has resulted in finer feature details on the face to be removed, or reduced.

3.2 De-noising

3.2.1 Smoothing

One of the simplest methods of reducing the amount of noise in images is to blur the pixels of the image so that noise is smoothed and leaves the main image features behind. This method also has the effect of softening hard edges within an image, therefore reducing details.

The 2D Gaussian function is a low-pass filter that attenuates the high frequencies and is commonly used for smoothing images. It is defined as

$$G_{2D}(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.16}$$

where x is the value in the horizontal axis and y is the value in the vertical axis. In statistics, the σ value refers to the SD of the Gaussian probability density function, and the square of it, σ^2, the variance. For the purposes of this work, the σ value refers to width of the Gaussian kernel. Fig. 3.4 shows an image that has been smoothed, or blurred, causing some details to be removed.

3.2.2 Temporal Noise Reduction

All videos captured can contain some form of noise due to the way images are captured digitally, whether this be through lighting, temperature or equipment malfunction. High-speed video is particularly susceptible due to capturing lots of images in a short space of time. To counter this, de-noising videos can be achieved using a sparse signal processing method [139] named collaborative filtering. This processes the video volume block-wise and attenuates noise to reveal fine details shared by the 3D block groups while preserving the unique features of each block. This method is named the Video Block Matching and 3D filtering (VBM3D). The output is the same size frames of the video, so no further processing is required to suit the feature descriptors.

The method uses a multitude of patches in the 3D neighbourhood of each pixel for attenuating the noise. The most similar neighbourhood patches are collected and stacked into a 3D array, where a 3D wavelet transform is applied and collaborative hard-thresholding is used for noise suppression. After the inverse transform is applied, the patches are returned to their original locations, and the average is taken. A second iteration is then completed with a Wiener filter to improve denoising results.

The overall algorithm can be described as follows. A noisy video $z(x) = y(x) + \eta(x)$, where y is the true video signal, $\eta(\cdot)$ $\mathcal{N}(0, \sigma^2)$ is an independent and identically distributed Gaussian noise sample and $x = (x_1, x_2, t) \in X$ are the coordinates in the spatio-temporal 3D domain $X \subset \mathbb{Z}^3$. The first two components $(x_1, x_2) \in \mathbb{Z}^2$ are the spatial coordinates and the third, denoted by $t \in \mathbb{Z}$, is the frame index (or time). The variance is assumed to be a priori and denoted by σ^2. This method described by Dabov et al. [139] is a signal processing method to denoise videos, therefore also denoising over time rather than just spatially. A limitation of this method is the computational complexity, meaning the processing times can be high if the videos are large or perhaps recorded at different frame rates. The authors provide a partial solution by adjusting parameters of the sliding step, N_{step} and the block-matching parameters N_S, N_{PR}, and N_B.

3.3 Feature Descriptors

3.3.1 Local Binary Patterns

The Local Binary Patterns (LBP) operator forms labels for each pixel in an image by thresholding a 3×3 neighbourhood of each pixel with the centre value. The result is a binary number where if the outside pixels are equal to or greater than the centre pixel, it is assigned a 1, otherwise it is assigned a 0. The amount of labels will therefore be $2^8 = 256$ labels.

This operator was extended to use neighbourhoods of different sizes. Using a circular neighbourhood and bilinearly interpolating values at non-integer pixel coordinates allow any radius and number of pixels in the neighborhood. The grey-scale variance of the local neighbourhood can be used as the complementary contrast method. The following notation of (P, R) will be used for pixel neighbourhoods, where P are sampling points on a circle of radius R. Fig. 3.6 shows an example of LBP computation.

Uniform patterns are used to reduce the length of the overall feature vector and implement a single rotation-invariant descriptor. An LBP that is uniform when the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. So 00000000 (0 transitions), 01110000 (2 transitions) and 11001111 (2 transitions) are uniform whereas the patterns 11001001 (4 transitions) and 01010010 (6 transitions) are not. In the computation of the LBP labels, uniform patterns are used so that there is a separate label for each uniform pattern and all the non-uniform patterns are labelled with a single label. For example, when using $(8, R)$ neighbourhood, there are a total of 256 patterns, 58 of which are uniform, which yields in 59 different labels.

Based on the LBP operator, Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) was first described as a texture descriptor [94] that used XT and YT temporal planes rather than just the 2D XY spatial plane. Yan et al. [17] used this method to report initial findings in the CASME II dataset, and Pfister et al. [64] use it as their feature descriptor. Fig. 3.5 shows a visualisation of an image with the LBP operator applied. Note that the image shown is for understanding how the features are applied and are not used in processing. This task is completed using the LBP histograms described earlier. Each region has the standard LBP

FIGURE 3.5: The original image is on the left. The LBP operator is applied and results in the right image. Notice that the face is still recognisable even when the image has been processed into binary patterns.

operator applied [140] with c being the centre pixel and P being neighbouring pixels with a radius of R

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{3.17}$$

where g_c is the grey value of the centre pixel and g_p is the grey value of the p-th neighbouring pixel around R. 2^p defines weights to neighbouring pixel locations and is used to obtain the decimal value. The sign function to determine what binary value is assigned to the pattern is calculated as

$$s(\mathbf{A}) = \begin{cases} 1, & \text{if } \mathbf{A} \geq 0 \\ 0, & \text{if } \mathbf{A} < 0 \end{cases} \tag{3.18}$$

If the grey value of P is larger than or equal to c, then the binary value is 1, otherwise it will be 0. Fig. 3.6 illustrates the sign function on a neighbourhood of pixels. After the image has been assigned LBPs, the histogram can be calculated by

$$H_i = \sum_{x,y} I\{LBP_l(x,y) = i\}, i = 0, \ldots, n-1 \tag{3.19}$$

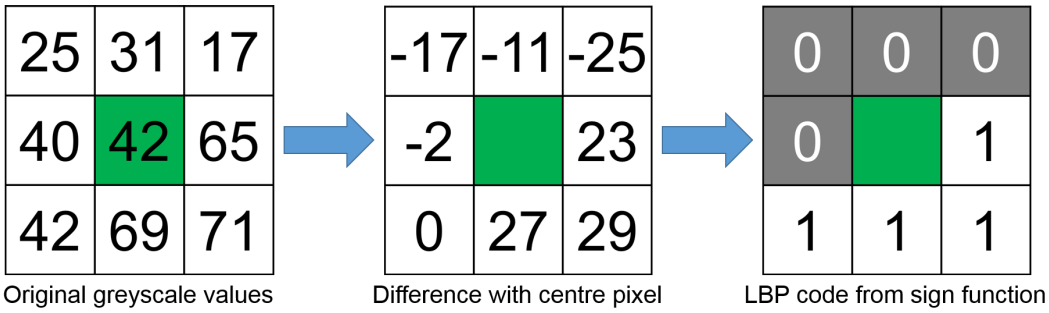

FIGURE 3.6: LBP code calculation by using the difference of the neighbourhood pixels around the centre.

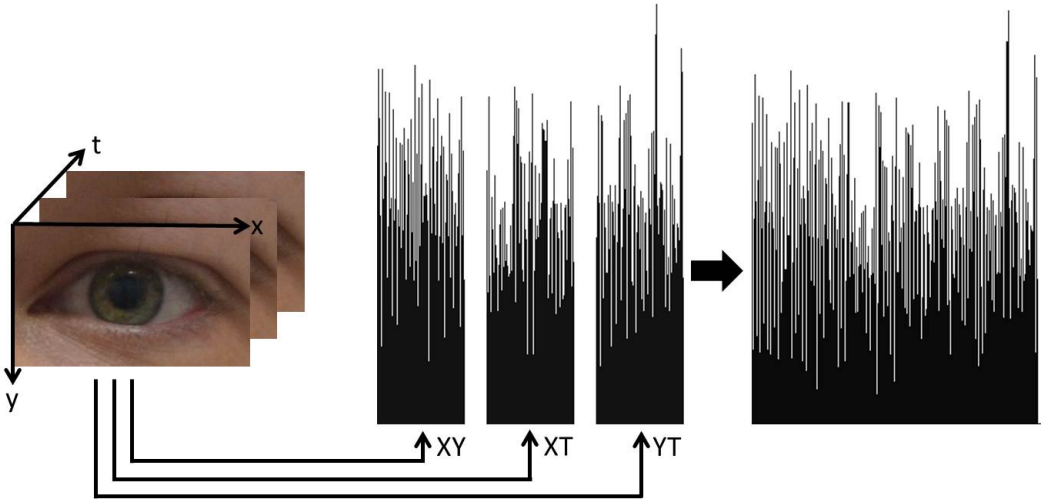

FIGURE 3.7: LBP is calculated on every block in all three planes. Each plane is then concatenated to obtain the final LBP-TOP feature histogram.

where $LBP_l(x, y)$ is the image labelled with LBPs. As this method is incorporating temporal data, the histogram can be extended to be calculated for all three planes

$$H_{i,j} = \sum_{x,y,t} I\{LBP_j(x, y, t) = i\}, i = 0, \ldots, n_j - 1 \tag{3.20}$$

where n_j is the number of labels produced by the LBP operator in the jth plane. $j = 0, 1, 2$ which represents the XY, XT and YT planes respectively. $LBP_i(x, y, t)$ expresses the LBP code of the central pixel (x, y, t) in the jth plane. The $I\{\mathbf{A}\}$ function is the equivalent to Eq. 3.19 that refers to the sign function in Eq. 3.18. An illustration of the LBP-TOP histogram concatenation process can be seen in Fig. 3.7.

The neighbouring points and radius parameters (P, R) can be defined as $P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T$ for each plane and axis, with the overall feature descriptor defined as $LBPTOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$.

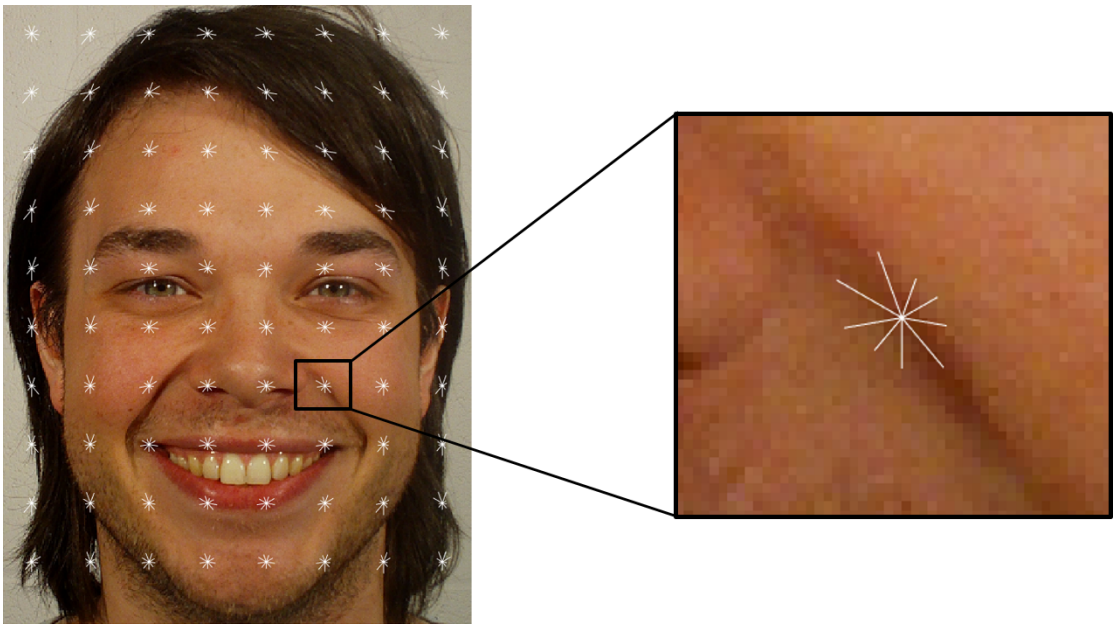

FIGURE 3.8: A visualisation of HOG on a face, where the small white plots represent a signed directional HOG cell weighted by the pixel magnitude. The zoomed in section shows a furrow on the face and how HOG represents it as a feature.

3.3.2 Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) features [11] were originally created for human detection in 2D images and used the pixel orientation values, weighted by its magnitude, to calculate features for describing a human as an object. Fig. 3.8 shows a visualisation of an image with the HOG operator applied. The image shown is for understanding how the features are applied and are not used in processing. The small white plots on the image denote the direction of a HOG cell weighted by the pixel magnitude using signed calculations. So, the longer the white line, the higher the magnitude in that direction. Each white line represents a particular bin, and in this example there are 9 bins in a 360 degree (or 2π) available orientations split into 40 degrees per bin. Polikovsky et al. [5, 98] used a 3D gradient histogram descriptor to recognise micro-expressions from high-speed videos. The paper used manually marked up areas that are relevant to FACS [39] based movement so that unnecessary parts of the face are left out. This does means that the method of classifying movement in these subjectively selected areas is time-consuming and would not suit a real-time application like interrogation.

The spatio-temporal domain is explored highlighting the importance of the temporal plane in micro-expressions, however the bin selection for the XY plane

is 8 and the XT, YT planes have been set to 12. The 8 bins in the XY planes was said to represent the different directions of movement, however the authors do not justify the temporal bin selection. An improvement on this would be to gradually increase bin size to find the ideal size for this purpose using 3D gradient descriptors.

The 3D Histogram of Oriented Gradients (3D HOG) descriptor has been adapted from Polikovsky et al. [5, 98] and like LBP-TOP it uses 3 planes to describe the spatial and temporal features. Gradients are calculated in the 3 dimensions of a video and the pixel orientation and magnitude is calculated for each plane. The magnitude values are binned into orientations so that the values are weighted based on the orientation of the gradient. Magnitude and orientation are defined as

$$Orientation(x, y) = \arctan\left(\frac{G_y}{G_x}\right) \tag{3.21}$$

$$Magnitude(x, y) = \sqrt{(G_x)^2 + (G_y)^2} \tag{3.22}$$

where G_x and G_y are the gradients of the x and y spatial directions respectively. The original HOG descriptor is then applied to each plane using Dollar's Matlab toolbox using the implementation described in [141, 142]. In contrast to the original HOG descriptor, the orientation defined for this method is 2π instead of π. Pixel orientation (Eq. 3.21) and magnitude (Eq. 3.22) are calculated and the magnitude values are binned into particular orientations so that the values are weighted based on the orientation of the gradient. The histogram bin parameter selection was the same as in the Polikovsky et al. [5, 98] protocols, where XY uses 8 bins and XT and YT uses 12 bins with 2 'no change' bins. Each plane is then concatenated to form the final 32-bin feature descriptor.

3.3.3 Optical Flow

Optical flow is the apparent motion of objects within images between two frames. The motion can be caused by camera movement or within-image object motion. It assumes that the pixel intensities do not change between frames and neighbouring pixel same similar motion.

If $I(x, y, t)$ is a pixel in an image, the brightness constancy constraint is defined as

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{3.23}$$

where Δx, Δy, and Δt are the small movements between the two frames. Assuming the movement is small, the Taylor series of the image constraint can be developed by

$$\begin{gathered} I(x + \Delta x, y + \Delta y, t + \Delta t) = \\ I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \rho. \end{gathered} \tag{3.24}$$

where ρ is a higher-order infinitesimal. From the previous equations it can be worked out that

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \tag{3.25}$$

or

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t}\frac{\Delta t}{\Delta t} = 0 \tag{3.26}$$

which results in

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \tag{3.27}$$

where V_x, V_y are the x and y components of the velocity respectively or the optical flow of $I(x, y, t)$. The derivatives of the image at (x, y, t) are $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$ and $\frac{\partial I}{\partial t}$ in the corresponding directions. This equation can be written more commonly as

$$I_x u + I_y v + I_t = 0 \tag{3.28}$$

where I is the image brightness and u and v are the horizontal and vertical optical flow vectors respectively.

The above calculations are not able to calculate optical flow vectors and contains two unknowns, u and v, that introduce the aperture problem. To solve this, further equations and constraints are required to compute the flow field.

3.3.3.1 Horn-Schunck

One of the first, and still widely used methods of computing optical flow vectors is the Horn-Schunk method [143]. It uses two main constraints, the first being

1/12	1/6	1/12
1/6	-1	1/6
1/12	1/6	1/12

FIGURE 3.9: The averaging kernel used in the Horn-Schunck method for iteratively computing the local averages of vectors.

the brightness constancy, the second was the smoothness constraint, that assumes adjacent pixels should move together as much as possible.

Also known as a regularisation equation, the Horn-Schunck method aims to minimise the following global energy function

$$E = \iint (I_x u + I_y v + I_t)^2 dxdy + \alpha \iint \left\{ \left(\frac{\delta u}{\delta x}\right)^2 + \left(\frac{\delta u}{\delta y}\right)^2 + \left(\frac{\delta v}{\delta x}\right)^2 + \left(\frac{\delta v}{\delta y}\right)^2 \right\} dxdy \tag{3.29}$$

where $\frac{\delta u}{\delta x}$ and $\frac{\delta u}{\delta y}$ are the spatial derivatives of the optical flow component, u, or the horizontal direction. The vertical component, v, is similar. To control the smoothness over the entire image, α can be scaled accordingly.

To minimise and obtain a optical flow field for each pixel in the image, the following equations are given which computes new velocity estimates from estimated derivatives and the average of previous velocity estimates

$$u_{x,y}^{k+1} = \bar{u}_{x,y}^{k} - \frac{I_x[I_x \bar{u}_{x,y}^{k} + I_y \bar{v}_{x,y}^{k} + I_t]}{\alpha^2 + I_x^2 + I_y^2} \tag{3.30}$$

$$v_{x,y}^{k+1} = \bar{v}_{x,y}^{k} - \frac{I_y[I_x \bar{u}_{x,y}^{k} + I_y \bar{v}_{x,y}^{k} + I_t]}{\alpha^2 + I_x^2 + I_y^2} \tag{3.31}$$

where $\begin{bmatrix} u_{x,y}^{k} & v_{x,y}^{k} \end{bmatrix}$ is the estimate of velocity for a pixel at (x, y) and $\begin{bmatrix} \bar{u}_{x,y}^{k} & \bar{v}_{x,y}^{k} \end{bmatrix}$ is the neighbourhood average. The original weighted average kernel used a 3×3 neighbourhood to find the averages of u and v. The kernel used can be seen in Fig.3.9.

3.3.3.2 Lucas-Kanade

In contrast to the global method of Horn-Schunck, the Lucas-Kanade method [144] attempts to solve u and v by dividing the original image into smaller sections and assumes a constant velocity in each section. A weighted least-squares fit of the brightness constancy equation is calculated for $\begin{bmatrix} u & v \end{bmatrix}^T$ in each section Ω. To complete this the following must be minimised

$$\sum_{x \in \Omega} W^2 [I_x u + I_y v + I_t]^2 \tag{3.32}$$

where W is a window function, which size can be set according to the requirements of the user. The minimisation problem is solved by

$$\begin{bmatrix} \sum W^2 I_x^2 & \sum W^2 I_x I_y \\ \sum W^2 I_y I_x & \sum W^2 I_y^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum W^2 I_x I_t \\ \sum W^2 I_y I_t \end{bmatrix}. \tag{3.33}$$

3.3.4 Histogram of Oriented Optical Flow

One of the most common ways of describing motion in images is to use optical flow. With the success of histogram of features, such as HOG in person recognition, Chaudhry et al. [145] proposed Histogram of Oriented Optical Flow (HOOF) features that bins optical flow vectors into histogram distributions. This approach is in contrast to the raw optical flow vectors, which may be of less use due to the number of pixels in a frame that can change over time. Moreover, optical flow computations are very susceptible to background noise, scale changes and directionality of movement.

To extract the HOOF features, we use the code provided by Uijlings et al. [146] but do not use the Bag-of-Words pipeline described. Instead we use parts of the code to replicate the original HOOF descriptor described in [145]. The flow vectors HOOF derives are calculated using the Horn-Schunck [143] approach that uses the brightness constancy assumption and global smoothness constraint that assumes the pixels moving in the image have smooth transitions in a pixel neighbourhood.

As with HOG, the flow vectors are binned into orientation bins based on calculated pixel angle and weighted by the magnitude. All optical flow vectors,

$v = [x, y]^T$ with direction, $\theta \tan^{-1}(\frac{y}{x})$ in the range

$$-\frac{\pi}{2} + \pi\frac{b-1}{B} \leq \theta < -\frac{\pi}{2} + \pi\frac{b}{B} \tag{3.34}$$

will contribute $\sqrt{x^2 + y^2}$ to the sum in bin b, $1 \leq b \leq B$, out of a total of B bins. The final histogram is then normalised to sum up to 1. Unlike LBP-TOP and 3D HOG, HOOF does not use three orthogonal planes as the flow vectors are already in the temporal domain.

3.3.5 Optical Strain

Optical flow calculates the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer (an eye or a camera) and the scene. Optical strain [6] is calculated from the optical flow vectors (using the Horn-Schunck [143] implementation) by taking the 2nd derivative in both the horizontal and vertical flow fields.

Optical strain was created as a way of detecting both macro- and micro-facial movements, the latter of which will be focused on here. It should be noted that optical strain in this thesis was reimplemented from the method proposed by Shreve et al. [6] and is not proposed by this research.

The horizontal and vertical motion vectors are represented by $\mathbf{p} = [p = dx/dt, q = dy/dt]^T$, with p being the horizontal vectors and q the vertical. Due to the smoothness or brightness constraint of optical flow, large intervals over a single expression can cause issues with tracking. To solve this, a vector stitching process was proposed that combines small intervals of around 1-3 frames into larger pairs to expand over the entire video sequence. As can be seen in Fig. 3.10, the stitching method works by matching the optical flow from one pair of frames, $\mathbf{p}(F_n, F_{n+1})$, to a consecutive pair of frames, $\mathbf{p}(F_{n+1}, F_{n+2})$. The (p, q) components are then summed to form the larger displacement $\mathbf{p}(F_n, F_{n+2})$.

The projected 2D displacement of an object that can be deformable is able to be expressed by a vector $\mathbf{u} = [u, v]^T$. For micro-facial movements, the motion tends to be small, so a finite strain tensor is defined as

$$\varepsilon = \frac{1}{2}[\nabla\mathbf{u} + (\nabla\mathbf{u})^T] \tag{3.35}$$

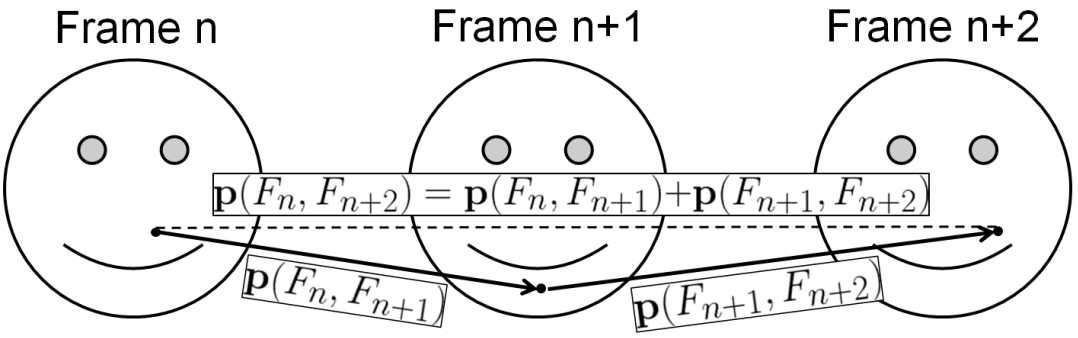

FIGURE 3.10: Vector stitching is applied to combine small intervals of optical flow vectors and create a larger displacement. Image reproduced from [6].

which can be expanded to

$$\varepsilon = \begin{bmatrix} \varepsilon_{xx} = \frac{\partial u}{\partial x} & \varepsilon_{xy} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \\ \varepsilon_{yx} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) & \varepsilon_{yy} = \frac{\partial v}{\partial y} \end{bmatrix} \tag{3.36}$$

where $(\varepsilon_{xx}, \varepsilon_{yy})$ are the normal strain components and $(\varepsilon_{xy}, \varepsilon_{yx})$ are the shear strain components. Each strain component defined is a function of displacement vectors (u, v) over a continuous space. They can be approximated using the discrete optical flow data (p, q) such that

$$p = \frac{\delta x}{\delta t} = \frac{\Delta x}{\Delta t} = \frac{u}{\Delta t}, u = p\Delta t \tag{3.37}$$

$$q = \frac{\delta y}{\delta t} = \frac{\Delta y}{\Delta t} = \frac{v}{\Delta t}, v = q\Delta t \tag{3.38}$$

where the change in time between two image frames is denoted as Δt. For this method, the Δt interval was set to a fixed length of 1 in order to maximise the sensitivity to micro-facial movements.

The partial derivatives of Eqs. 3.37 and 3.38 can be estimated in both the horizontal and vertical direction as follows

$$\frac{\partial u}{\partial x} = \frac{\partial p}{\partial x}\Delta t, \frac{\partial u}{\partial y} = \frac{\partial p}{\partial y}\Delta t \tag{3.39}$$

$$\frac{\partial v}{\partial x} = \frac{\partial q}{\partial x}\Delta t, \frac{\partial v}{\partial y} = \frac{\partial q}{\partial y}\Delta t. \tag{3.40}$$

The strain components are finally estimated using the central difference method by convolving two 3×3 Sobel kernels over each of the (p, q) flow fields to generate each of the four strain components of the strain tensor defined in Eq. 3.35. The

first Sobel kernel (S_x) that estimates the horizontal derivatives is defined as

$$S_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{3.41}$$

and the second Sobel kernel (S_y) estimates the vertical derivatives and is defined as

$$S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}. \tag{3.42}$$

The central difference method used to calculate the second order derivatives is calculated by

$$\frac{\partial u}{\partial x} = \frac{u(x+\Delta x) - u(x-\Delta x)}{2\Delta x} = \frac{p(x+\Delta x) - p(x-\Delta x)}{2\Delta x} \tag{3.43}$$

$$\frac{\partial v}{\partial y} = \frac{v(y+\Delta y) - v(y-\Delta y)}{2\Delta y} = \frac{q(y+\Delta y) - q(y-\Delta y)}{2\Delta y} \tag{3.44}$$

where $(\Delta x, \Delta y)$ is a change of 1 pixel. All equations have been replicated from [6, 93, 107].

Finally, after the calculation of the strain patterns, a visual representation, called a strain map, can be created by normalising the strain magnitude values to 0-255. The benefit of creating this visualisation is that the estimated motion on the face can be highlighted and is shown in Fig. 3.11.

3.4 Methods of Classification

Machine learning is all about learning structure from data, and many methods exist in this field. In this section, two classification methods, SVM and RF, are described.

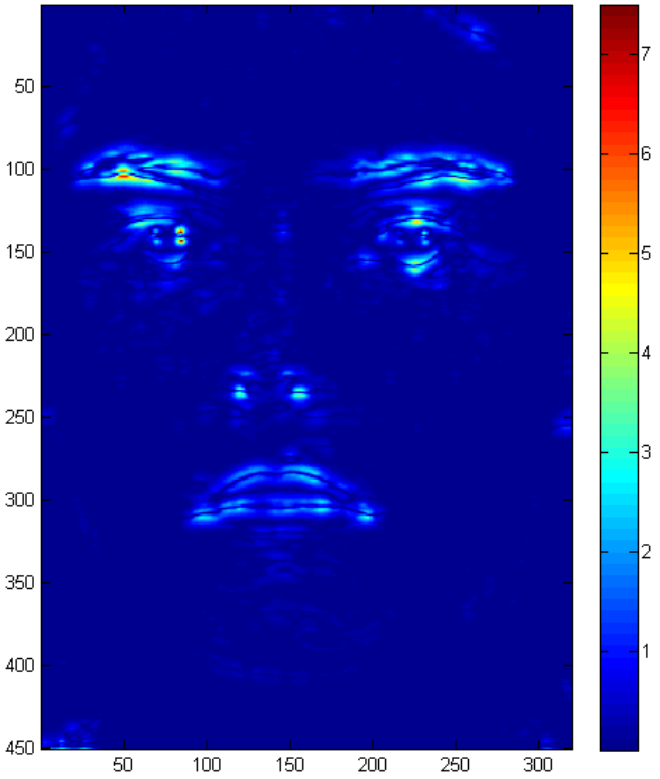

FIGURE 3.11: A strain map calculated from the normalised optical strain magnitudes. The motion on areas of the face are estimated in the colours closer to red.

3.4.1 Support Vector Machines

First proposed by Cortes and Vapnik [12] an SVM attempts to find a linear decision surface (hyperplane) that can separate classes and has the largest distance between support vectors (elements in data closest to each other across classes). If a linear surface does not exist, then an SVM is able to use kernel functions to map the data into a higher dimensional space where a decision surface can be found. SVM was originally based on the Structural Risk Minimisation principle, which was used for machine learning from a finite dataset.

As shown in Fig. 3.12, data points are split using an optimal separating hyperplane. The dashed lines on either side of the hyperplane is hereby defined as the margin m. Each training vector $\mathbf{x}$ belongs to a class y, with the training set defined as $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. The total set and classes are defined as $(\mathbf{x}_i) \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ where $\mathbb{R}^d$ is a real number in d-dimensions and $\{-1, +1\}$ are the two classes. For a given hyperplane, $\mathbf{x}_+$ and $\mathbf{x}_-$ are the closest points to the hyperplane among the positive and negative examples. The norm of a vector $\mathbf{w}$

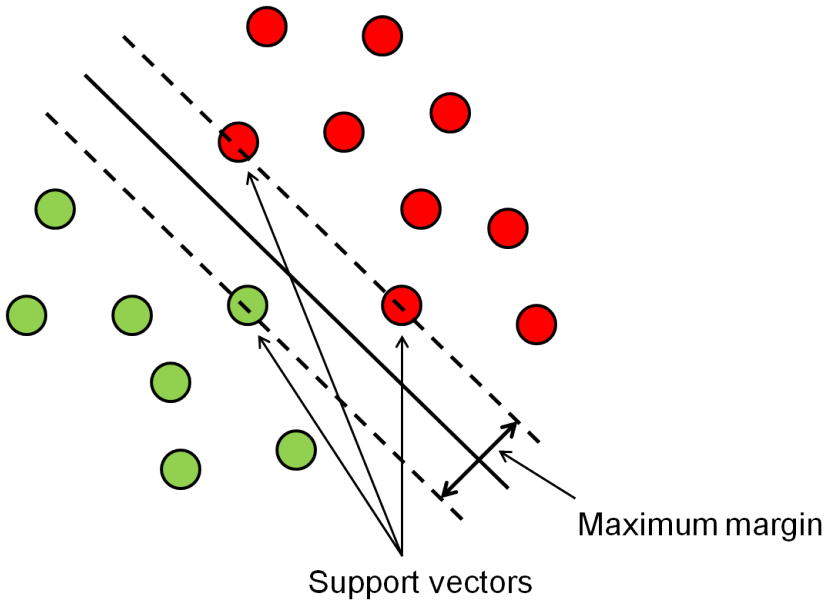

FIGURE 3.12: Visualisation of an SVM hyperplane. The green and red circles represent the positive and negative classes respectively, with the support vectors contributing to hyperplane separation leading to the determination of the maximum margin.

is denoted by $||\mathbf{w}||$ as its length and is given by $\sqrt{\mathbf{w}^T\mathbf{w}}$. A unit vector $\mathbf{w}$ in the direction of $\mathbf{w}$ is given by $\mathbf{w}/||\mathbf{w}||$ and $||\mathbf{w}|| = 1$.

From a geometric consideration, the margin of a hyperplane h with respect to a dataset D can be defined as

$$m_D(f) = \frac{1}{2}\mathbf{w}^T(\mathbf{w}_+ - \mathbf{w}_-) \tag{3.45}$$

where there is an assumptions that $\mathbf{w}_+$ and $\mathbf{w}_-$) are equidistant from the decision boundary as

$$f(\mathbf{x}_+) = \mathbf{w}^T\mathbf{x}_+ + b = a \tag{3.46}$$

$$f(\mathbf{x}_-) = \mathbf{w}^T\mathbf{x}_- + b = -a \tag{3.47}$$

for some constant $a > 0$. To make this geometric margin meaningful, the value of the decision for the points closest to the hyperplane, $a = 1$. By adding Eq. 3.46 and Eq. 3.47 and then dividing by $||\mathbf{w}||$, the margin becomes

$$m_D(f) = \frac{1}{2}\mathbf{w}^T(\mathbf{w}_+ - \mathbf{w}_-) = \frac{1}{||\mathbf{w}||} \tag{3.48}$$

Next, a maximum margin classifier, sometimes called a hard margin, is defined

to handle linearly separable data. It can then be modified to attempt to handle less easily separable (or non-separable) data. The maximum margin classifier is the discriminant function that maximises the geometric margin $1/||\mathbf{w}||$ which is the equivalent to minimising $||\mathbf{w}^2||$. This leads to the following constrained optimisation problem

$$\begin{aligned} \underset{\mathbf{x},b}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{w}^2|| \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, 2, \ldots, n \end{aligned} \tag{3.49}$$

where the constraints show ensure that the maximum margin classifies each example correctly assuming the data is linearly separable. However, it is often the case that data is not linearly separable. A larger margin can be determined by allowing for some misclassification of points. The optimisation problem now becomes

$$\begin{aligned} \underset{\mathbf{x},b}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{w}^2|| + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{3.50}$$

where $\xi \geq 0$ are the variables that allow for a margin error, $0 \leq \xi_i \leq 1$, or to be misclassified by $\xi > 1$. The constant $C > 0$ sets the relative importance of maximising the margin and minimising the amount of errors. This way of calculating for non-separable data is called a soft margin SVM.

Lagrange multipliers are used as a mathematical method to solve constrained optimisation problems of differentiable functions. With an SVM, the saddle point of the Lagrange function can be found using

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}||\mathbf{w}^2|| - \sum_{i=1}^{n}\alpha_i\{y_i[(\mathbf{w}^T \cdot \mathbf{x}_i) + b] - 1\} \tag{3.51}$$

where α_i are the Lagrange multipliers. The Lagrangian function has to be minimised with respect to $\mathbf{w}, b$ and maximised with respect to $\alpha_i \geq 0$. The optimisation can be transformed into its dual problem as

$$\max_{\alpha} D(\alpha) = \max_{\alpha}\left\{\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)\right\} \tag{3.52}$$

and the optimal separating hyperplane is represented by the dual solution

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \cdot y_i \cdot \mathbf{x}_i \tag{3.53}$$

The value of b can be estimated by inputting $\mathbf{w}$ into the original equation $\mathbf{w}^T\mathbf{x} + b = 0$. For testing, the classification is given by

$$f(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b) \tag{3.54}$$

for any new data point $\mathbf{x}$. If the training data input into the SVM is non-separable, then the error variables, ξ, can be used.

For this thesis, the kernel used for the SVM was the Radial Basis function. During empirical experiments, the cost parameter was found to be optimal set to 1, gamma to 1/number of features and ϵ to 0.001. LibSVM within Weka was used as the SVM classifier.

3.4.2 Random Forests

A relatively new machine learning approach, developed by Breiman [26], has the idea that if one classification tree is good, then many trees (a forest) should be better. Random Forests (RF) can run efficiently on large datasets, with only the storage requirements of the dataset being the major memory requirements. They are also resistant to overfitting (i.e. the model described is random error or noise instead of the underlying relationship), therefore the performance of this algorithm does not decrease as the number of trees increases.

Firstly, each tree is trained on around two thirds of the training data provided to the algorithm, with each case picked at random from the original data. Selection of this data is also named bootstrap aggregation, which randomly generates a number of decision trees (denoted as n_{tree}), which are each provided with randomly selected samples of the training input and then all decision trees are combined into a decision forest. Then, some predictor variables, m, are randomly selected from all of the predictor variables to best split the node. By default, m is set to the square root of the total number of predictors for classification and stays constant during the growing of the forest. By changing m (change denoted by m_{try}), the

RF can be tuned for different data. The number of trees used in a forest, n_{tree}, can also be adjusted to fine tune parameters in this method.

The remaining one third of data left from the training set is used to calculate the misclassification rate named the out of bag (OOB) error rate. The combined error from all trees is used to determine the overall OOB error rate for classification. The error rate of a forest can depend on two main points. First, the correlations between any two trees in a forest, therefore increasing the correlation increases the error rate. Second, trees with low error rates are classed as strong classifiers, and so increasing the strength of individual trees will decrease the overall forest error rate. Reducing m_{try} reduces both the correlation and strength, and vice versa.

Finally, each tree provides a classification choice, and it can be said that the tree has voted for that particular class. Whichever classification has the most votes from all the tree is chosen as the correct class. For example, in a binary classification problem, the vote would be in either yes or no, with the RF score is the percentage of the yes votes and is the predicted probability.

3.5 Feature Difference Measures

3.5.1 Sum of Squared Differences

One of the simplest ways of finding the difference between some points, is to take the sum of squares. For the use in micro-movement analysis, it can be used to find the difference between frames in the micro-movement sequence. The method works for a 1D vector of frame values, or more likely, histogram features that describe each frame, for example HOG or LBP.

There are two versions of the difference method. The first finds the difference between the first frame and the i-th frame. It is defined as

$$\sum_{i=1}^{n}(x_1 - x_i)^2 \tag{3.55}$$

where x_i is the frame in the video sequence. This equation is similar to using the first frame as a reference frame. The second version calculates the difference

between the i-th frame and the i-th - 1 frame. It is defined as

$$\sum_{i=2}^{n}(x_i - x_{i-1})^2 \tag{3.56}$$

where in contrast to the previous equation, compares the current frame with the previous frame as it iterates.

3.5.2 Chi-Square Distance

The concept of feature difference analysis for micro-movement analysis was introduced by Moilanen et al. [109]. For each current frame being processed, it is compared with the average feature frame that is represented by the average features of the start frame, k-th frame before the current frame, and the end frame, k-th frame after the current frame. The k-th frame is described as

$$k = \frac{1}{2}(N - 1) \tag{3.57}$$

where N is the micro-interval value that is always set to an odd number. Moilanen et al. [109] sets the micro-interval to 9 for the SMIC dataset [15] recorded at 25 fps and 21 for the CASME [16] dataset that was recorded at 60 fps.

The difference between the current frame and average feature frame shows the facial changes in a particular region and the possible change in the features is rapid since it occurs between start frame and end frame, to distinguish the quick changes from temporally longer events.

The difference analysis continues for each frame except the first and last k frames that would exceed the boundaries of the video. This also means rapid facial movements such as blinks would also be classified as a movement.

The difference that is calculate can also be described as the dissimilarity between histograms in each region. The χ^2 distance is such as measure and is defined as

$$\chi^2(P, Q) = \sum_i \frac{(P_i - Q_i)^2}{(P_i + Q_i)} \tag{3.58}$$

where i is the i-th bin in the P and Q histograms that have an equal number of bins. This equation can be used in all temporal planes (XY, XT, YT) to calculate dissimilarity.

In many other histogram distance measures, the differences between large bin values are less significant than between small bins [147], however as this is calculating dissimilarity the difference should not be dependent on the scale of the bin values when each feature is deemed equally important. Ahonen et al. [148] found that the χ^2 distance performs better in face analysis task than histogram intersection or log-likelihood distance. It has also been applied successfully to applications such as object and text classification [149].

With the feature differences calculated for each block and frame of a video sequence, an initial feature vector is produced by taking the average of the M greatest block difference values. Moilanen et al. [109] chose 12 as the value of M which corresponds to the 12 blocks out of a 26 block total. The initial feature vector is designed to represent the whole video and is defined as

$$F = \frac{1}{M}\sum_{j=1}^{M}(D_{1,j}, D_{2,j}, \ldots, D_{n,j}) \tag{3.59}$$

where D is an $(n \times b)$ corresponding to the b block difference values that are sorted in descending order for each frame, and n is the total number of frames. In this method the matrix would be $(n \times 36)$.

It was determined in [109] that by setting M to around one third of the block difference values would give better discrimination than one maximum value or an average of all block values, for example. This conclusion of one third did not go into detail, and more information would be required to assume this is correct.

To reduce noise and local magnitude variations, a contrasting feature vectors was created by subtracting the average of the surrounding tail frame and head frame values from each current frame value in F. Therefore, the i-th value in the contrasted feature vector is defined as

$$C_i = F_i - \frac{1}{2}(F_{i+k} + F_{i-k}) \tag{3.60}$$

where C_i and F_i are the contrasted feature vectors and initial feature vector at the i-th value respectively. The completed contrasted feature vector being calculated for the whole video using Eq. 3.60 for all n frames. Due to the temporal nature of this method, the first and last k frames cannot be included in the calculation.

After C is created, all the negative values are found and assigned to be zero as these denote no rapid changes of the current frame compared to the tail frame and head frame.

Finally, a threshold is calculated that is used as a single value to determine at what point a peak should cross to be counted as a micro-facial movement. This is calculated as

$$T = C_{mean} + p \times (C_{max} - C_{mean}) \tag{3.61}$$

where C_{max} and C_{mean} are the maximum and average values for the whole video, and p is a percentage parameter in a range of 0 - 1. Peak detection was reportedly used, however with no explicit mention of how it was done, there is a chance the peak annotation was manual.

3.5.3 Peak Detection

To detect peaks automatically from feature difference vectors, a signal processing method can be applied [150]. The first derivative of the signal is smoothed, and then peaks are found from the downward-going zero-crossings. Then only the zero-crossings whose slope exceeds a defined 'slope threshold' at a point where the original signal exceeds a certain height or 'amplitude threshold'.

3.5.3.1 Temporal Phases of Peaks

Estimating the apex simply uses the highest point of the found Gaussian curves. Estimating the 'start' and 'end' of the peak (the onset and offset value respectively), is a bit more arbitrary, because typical peak shapes approach the baseline asymptotically far from the peak maximum. Peak start and end points can be set to around 1% of the peak height, but any random noise on the points during peak detection will often be a large fraction of the signal amplitude at that point.

Smoothing is done within this process to reduce noise, however this can lead to distortion of peak shapes and change the start and end points. One solution is to fit each peak to a model shape, then calculate the peak start and end from the model expression. This minimises the noise problem by fitting the data over the entire peak, but it works only if the peaks can be accurately modelled. For

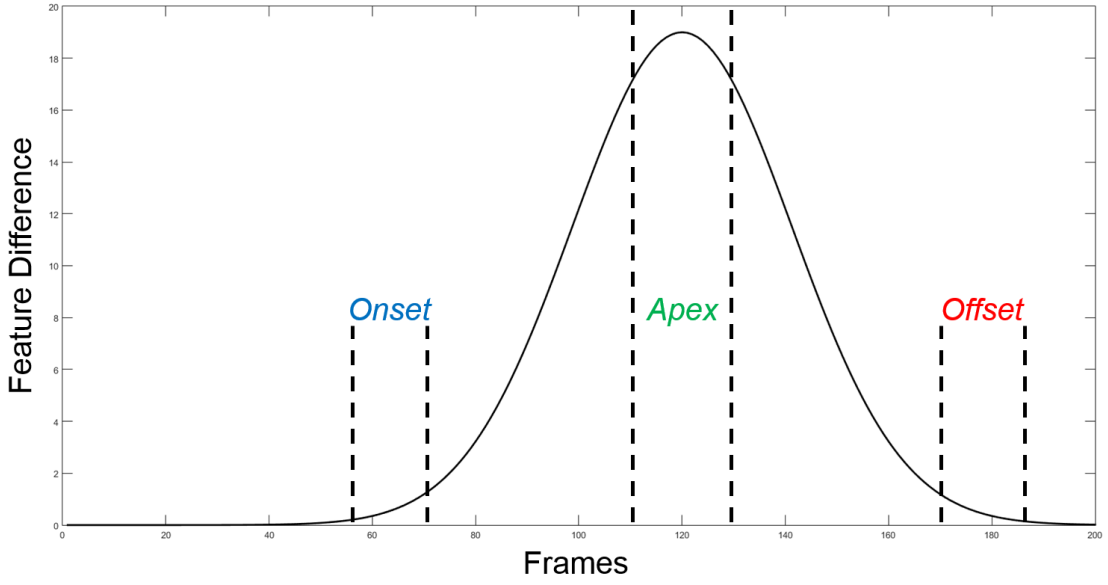

FIGURE 3.13: The temporal phases used to denote the onset, apex and offset of a micro-facial movements, facial expressions and other similarly shaped difference features.

example, Gaussian peaks reach a fraction a of the peak height of

$$x = p \pm \sqrt{\left(\frac{w^2 \log(\frac{1}{a})}{2\sqrt{(\log(2))}} \right)} \tag{3.62}$$

where p is the peak position and w is the peak width [150]. So if $a = 0.01$, $x = p \pm 1.288784 \times w$. An example of stereotypical Gaussian curve shape with added temporal phases used are shown in Fig. 3.13.

3.6 Performance Measures

To measure the performance of the micro-movement detection, quantification of results will be presented in four measurements: *Recall*, *Precision*, *F-Measure* and *Matthews Correlation Coefficient* (MCC). These measurements are commonly used for binary classification purposes, and so is adequate for quantifying True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) detections. More detailed information on these measures can be found in [151].

3.6.1 Precision and Recall

By using the *Precision* measure of exactness, and determines a fraction of relevant responses from results. Recall, or sensitivity, is a fraction of the results that are relevant to the experiment and that are successfully retrieved.

$$Precision = \frac{TP}{TP + FP} \tag{3.63}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.64}$$

It is unlikely to use these measures on their own as both these measure are commonly used together to form an understanding of the relevance of the results returned from experimental classification.

3.6.2 F-Measure

The F-Measure is useful in determining the harmonic mean between the *Precision* and *Recall* and is used in place of accuracy as it provides a more detailed analysis of the data. The equation can be defined as

$$F\text{-}Measure = \frac{2TP}{2TP + FP + FN}. \tag{3.65}$$

A downside to this measure is that it does not take into account TNs, a value that is required to create Receiver Operating Characteristic (ROC) curves.

3.6.3 Matthews Correlation Coefficient

The Matthew's Correlation Coefficient (MCC) uses all detection types to output a value between -1, which indicates total disagreement and $+1$, which indicates total agreement. A value of 0 would be classed as a random prediction, and therefore both variables can be deemed independent. It can be provide a much more balanced evaluation of prediction than previous measurements, however it is not always possible to obtain all four detection types (i.e. TP, FP, FN, TN). The

coefficient can be calculated by

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.66}$$

3.7 Summary

All of the theories and techniques described in this Chapter provides the foundation on which the following contribution Chapters will be based. Chapter 4 will describe preliminary studies into micro-movement detection using a new feature: LBP-TOP combined with GD. Chapter 5 introduces a new micro-movement detection corpus, SAMM, created through an emotional inducement experiment. Chapter 6 outlines a method of using human facial expression baseline analysis, combined with a feature difference method, to detect micro-movements. The final contribution in Chapter 7 will outline new FACS-based face regions, combined with spatio-temporal features, that will enable the localisation of micro-movement occurrence on the face.

Chapter 4

Micro-Movement Detection: Preliminary Studies

This Chapter presents preliminary investigations into recognising micro-facial movements, which form the basis of the extended LBP-TOP feature combined with Gaussian Derivatives (GD) for recognising micro-movements. The focus of this thesis is on micro-movements, however a thorough exploration into established methods is required for comparison.

4.1 Introduction

Recognising micro-expressions in early literature is commonly referred to as micro-facial expression recognition. The focus of this field is to categorise all micro-expressions into specific classes, for example, positive, negative and surprise [15]. Unlike macro-facial expressions, the subtle expressions cannot be clearly defined into distinct classes. In light of this, an investigation is done to recognise the movement and non-movements of the face. This method is named micro-movement recognition, and follows a more objective goal of determining what is a micro-movement, and what is a non-movement.

With the large amounts of research dedicated to using machine learning to recognise and classify facial-expressions into AUs or emotions groups, it is only

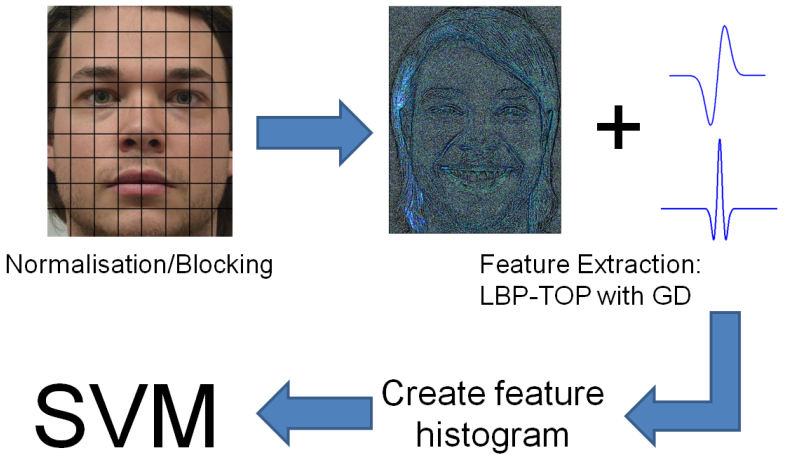

FIGURE 4.1: System summary of LBP-TOP with GD and classified with SVM.

logical to investigate if the already established methods would work effectively for micro-movement analysis.

After describing different features and approaches in Section 4.2, a new feature extending LBP-TOP with Gaussian Derivatives is proposed [152] in Section 4.3. LBP-TOP has been used both in the analysis of facial expressions [94] and micro-facial expressions [15, 17], and is known to represent the face well. To improve the extraction of more important facial features, LBP-TOP is combined with Gaussian Derivatives (GD) by calculating the first and second order GD on every image to highlight facial features. The LBP-TOP operator is then applied to the video sequence the to form the final LBP-TOP histogram features. The overall system summary can be seen in Fig. 4.1. Using RF and SVM machine learning algorithms, a training and testing set are created to find a model that will accurately classify a movement and a non-movement as separate. These algorithms were chosen for the popularity and good results obtained in facial analysis methods.

This Chapter will begin outlining investigations that will then inform on the creation of a novel feature and recognition tests. Completing these allows for a clearer understanding of the field and approach to experiments undertaken.

4.2 Preliminary Investigations

Prior to the main experiment, some initial tests with features already used in micro-expression or movement analysis were performed to obtain initial validation

results and gain an understanding of what the best representation feature was. All tests were performed on the CASME II [17] dataset and most used the Weka [153] data mining software to implement the SVM and RF learning algorithms. In addition, the face was split into 4×4 blocks and the features were calculated from these areas.

It should be noted that the CASME II dataset contains just over 200 micro-expressions, however the feature classification values described in this section are based on the histogram bin features, so the number of classification values will appear higher than expected. Each classification run was completed using 10-fold cross validation.

4.2.1 Optical Strain

The optical strain [6] was one of the features that could not be readily be put into the machine learning algorithms as the length of the video sequences change. If any normalisation was performed, the frame information would be lost and the onset, apex and offset would not be known. This information is a key aspect of using this type of feature.

After calculating on some test videos, the feature was found to be very sensitive to movements, so even the slightest head movement may influence the magnitude change in time. Good head registration or alignment is required to minimise rigid head motion while retaining the subtle facial muscle motions.

One of the better outcomes for optical strain was the visualisation of the strain patterns as shown in Fig. 4.2. The image can be interpreted as showing the strongest movement areas in red and weakest in blue. This particular strain pattern was taken from an AU7 movement around the eye, and the corresponding temporal plot of the strain magnitude over time can be seen in Fig. 4.3. The movement occurred around frame 18, however, due to the sensitive nature of this feature the peak cannot be easily distinguished. It was concluded that using optical strain produced no distinct patterns, and that further feature descriptors should be explored.

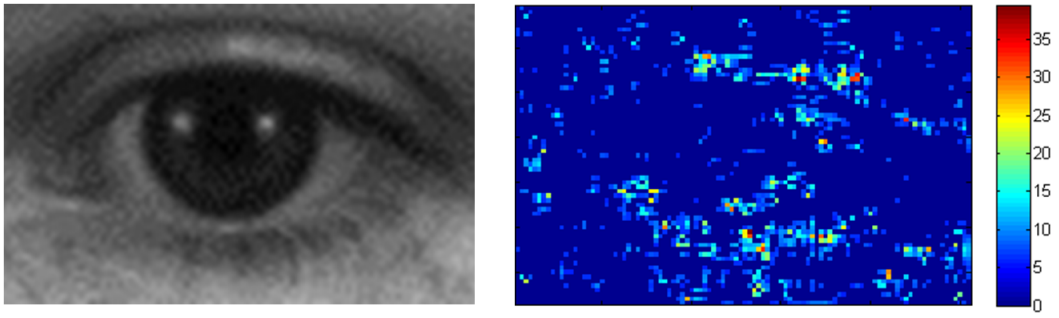

FIGURE 4.2: A still from the original eye movement is on the left, where AU7 occurs. The movement is when the inner eye muscle contracts and the eyelids slightly close. The optical strain pattern on the right describes where the highest level of movement occurs, with red being the strongest and blue the weakest.

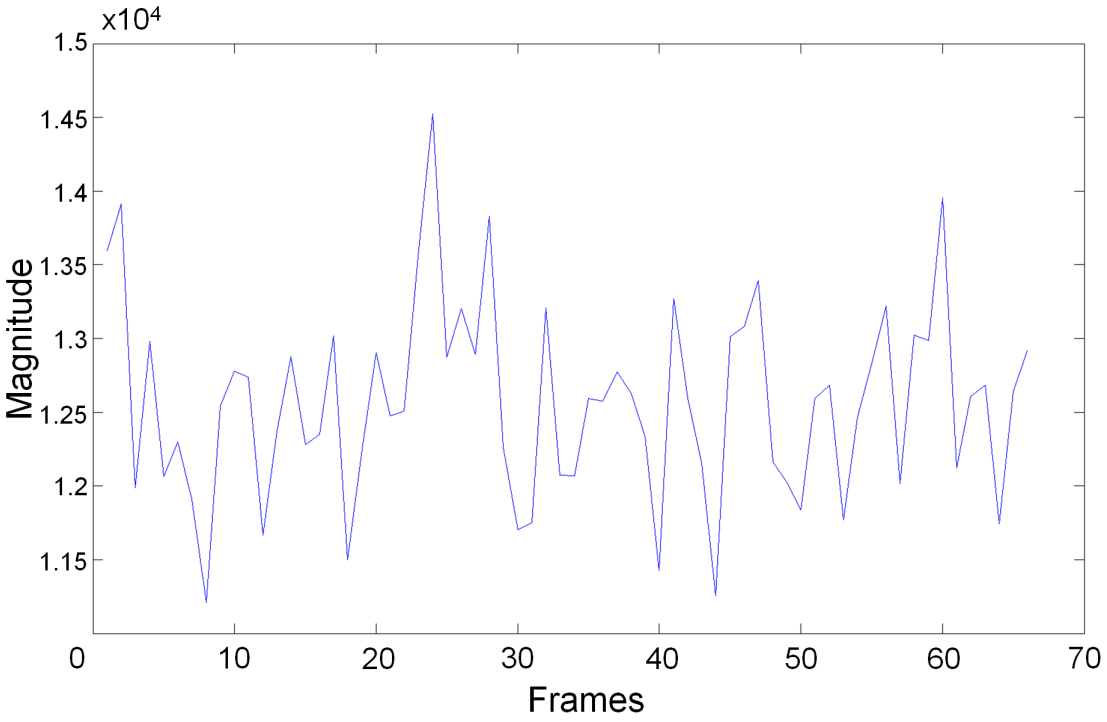

FIGURE 4.3: The temporal plot of the optical strain magnitude changes. The AU7 movement occurs around frame 18 and offsets around frame 30.

4.2.2 Feature Difference Using Sum of Squares

Another method that did not incorporate machine learning, and was only tested on a limited set from CASME II, was to use the sum of square differences (see Section 3.5.1). The two versions of sum of squares were used to describe the micro-movements. The first technique used compared all other frames with the first frame as a reference. Fig. 4.4 shows the result of processing the movement using this technique.

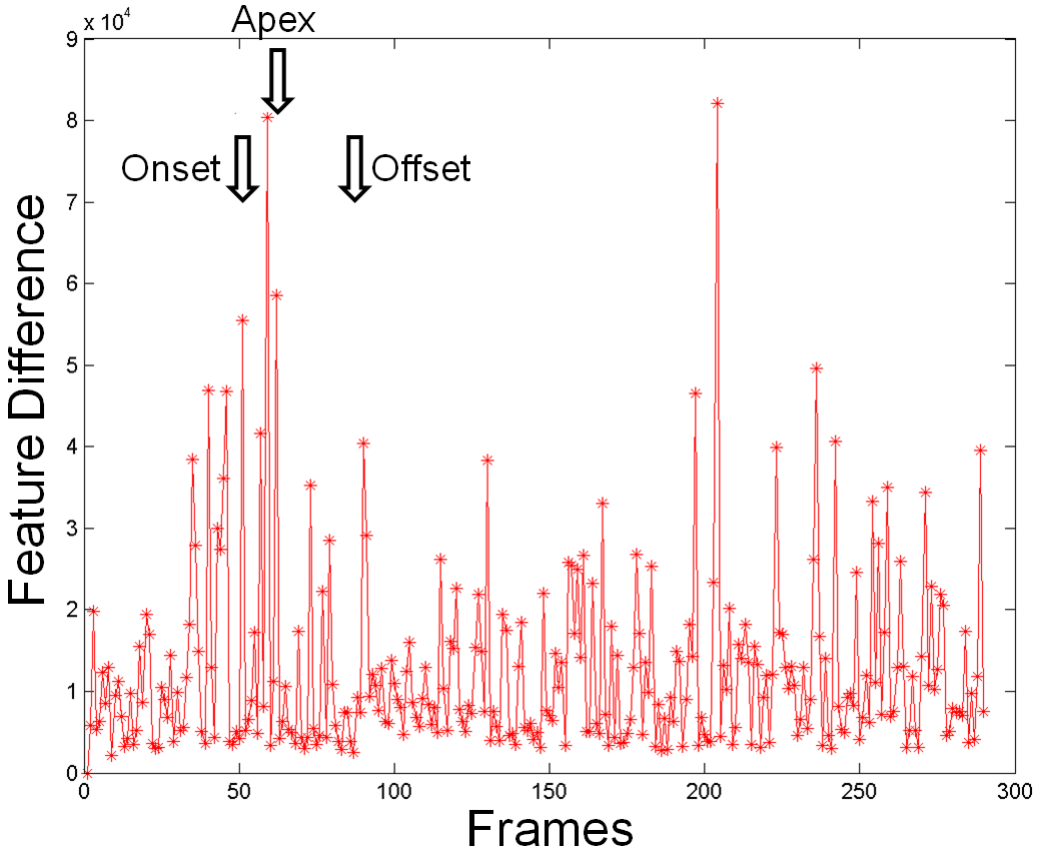

FIGURE 4.4: The result of the first version of the sum of square differences method on a micro-movement.

The same movement was then processed with the other technique, where the difference between the i-th frame and the i-th-1 frame is computed, therefore each frame was compared to the previous frame. The result of this technique is shown in Fig. 4.5.

As can be seen in both results, calculating the difference in this way can lead to a lot of noise that is not a movement. For clarity, the actual ground truth onset, apex and offset temporal phases have been included to show that the movement was not clearly defined as a peak. The majority of peaks are created by small pixel-based changes that the sum of square differences method is sensitive to.

Similarly to optical strain, it was concluded that, although a simple and reasonably quick method to implement, the features did not have distinct features to describe micro-movements. The features explored in the following sections better represent facial movement.

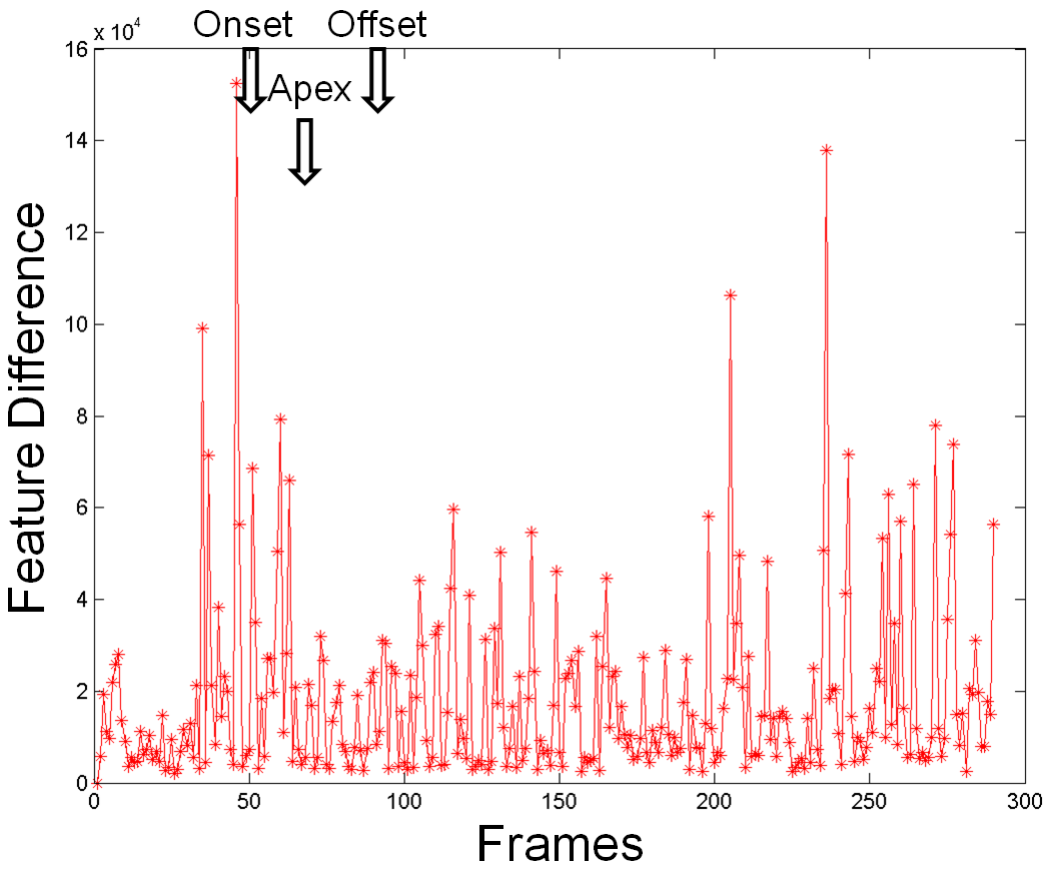

FIGURE 4.5: The result of the second version of the sum of square differences method on a micro-movement.

4.2.3 3D HOG

The temporal version of the HOG [11] descriptor was proposed by Polikovsky et al. [5] and uses the XY, XT and YT planes similar to LBP-TOP. The clipping value for this method was originally 0.2, however the XY spatial bins always summed higher than the other bins during preparation for a format the machine learning algorithms can understand (i.e. normalised feature to handle different video sequence frame lengths). In light of this, the clipping value of the 8 XY bins was set to 0.2, and an increased value of 1 for the two 12-bin temporal descriptors of XT and YT.

Firstly, the absolute distance between all frames in a sequence in each block is taken, and then resulting distances were summed. This attempted to find magnitude changes in sequences where non-movements would theoretically be low. The results in Weka were poor, with the distribution of values overlapping with no clear separation possibility.

Secondly, the minimum and maximum magnitude of each sequence was calculated to find the differences and use this as the final value. Again this was designed to split micro-movement and non-micro-movements as the movement magnitudes should have large differences between the minimum and maximum values whereas the non-movement should not. The results did no better than chance in either SVM or RF due to the lack of feature separation.

Finally a test on obtaining the median of each feature was completed, but the results were still poor in Weka and distribution of feature points was similar. The median appeared to have even greater weight to lower magnitudes, especially in the temporal planes.

After the above tests, it was concluded that no matter how the values are obtained for analysis, attempting to generalise the sequences into one value is difficult or potentially not viable. Micro-movements may require to keep as much information as possible to ensure a difference between micro-movements and non-movements can be found.

4.2.4 3D HOG with TIM

With 3D HOG not performing well alone, and with noisy results obtained from optical strain and sum of squares, a way of normalising the micro-movements using a Temporal Interpolation Model (TIM) [64] was used. TIM was used in conjunction with LBP-TOP [94] to form a feature that reduced the micro-movement sequences by interpolating to 10 frames (TIM10). A downside to this is it removes the temporal aspect of movements and so cannot identify the onset, apex and offset. However, it attempts to emphasise the micro-movements to reduce unnecessary movement frames and noise.

The first advantage for using 3D HOG with TIM10 was the descriptor calculated much faster due to the lower amount of frames to process. As with 3D HOG without TIM10, three ways of inputting the data into Weka were used. The first used the absolute distance between all frames in a sequence in each block, and then resulting distances were summed. The second found the minimum and maximum magnitude of each sequence was calculated to find the differences. Finally the median of the sequence was calculated. Unfortunately, even with the addition of TIM10, the results were no better than chance on all tests.

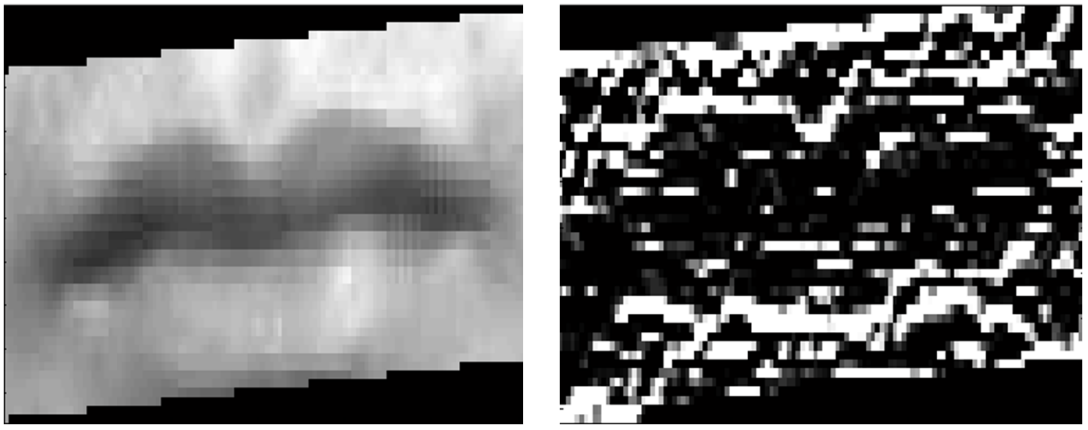

FIGURE 4.6: A visualisation of a micro-movement frame in the mouth region. The left hand image shows the original and the right hand shows the spatial XY plane.

As a feature, 3D HOG with TIM did not have to be normalised, however by reducing the amount of information available, including the direction and magnitudes of facial gradients, the recognition rates were below chance. Although 3D HOG appears to not be suitable as a feature for micro-movement analysis, the poor results are due to normalising the data to work with the machine learning software and algorithms. Further tests would be required that keep more feature information, such as the frame differences, to investigate this conclusion.

4.2.5 LBP-TOP with TIM

The final descriptor used was LBP-TOP with TIM10, previously implemented in [64]. Normalising the micro-movements to only 10 frames should focus on areas that are important and interpolate these effectively. A visualisation of the LBP-TOP feature in the XY plane, around the mouth, can be seen in Fig. 4.6.

All planes of the LBP-TOP descriptor were used for the following results as they highlight the best case outcome. Classification was unsuccessful with libSVM [154] in Weka due to the method not being able to find a separation between micro-movements and non-micro-movement classes. RF (using 100 trees) in Weka performed better, and the confusion matrix can be seen in Table 4.1.

The number of trees for RF was then changed to 70, with the performance of the feature increasing slightly. The confusion matrix can be seen in Table 4.2. Using this feature was promising, as it predicted a high number of micro-movements

TABLE 4.1: The confusion matrix of the micro-movement recognition results using RF with the amount of trees set to 100.

	Predicted Micro-Movements	Predicted Non-Micro-Movements
Actual Micro-Movements	750	497
Actual Non-Micro-Movements	220	2549

TABLE 4.2: The confusion matrix of the micro-movement recognition results using RF with the amount of trees set to 70.

	Predicted Micro-Movements	Predicted Non-Micro-Movements
Actual Micro-Movements	754	493
Actual Non-Micro-Movements	212	2557

when using RF, however SVM did not perform as well. Further, LBP-TOP was able to be normalised to a 177 dimension feature vector, making it very suitable for inputting the data into machine learning. However, using TIM10 creates an artificial interpolation of micro-movements, and keeping as much information as possible is advantageous.

Due to the results presented, the LBP-TOP descriptor is the most suitable as an initial feature in micro-movement recognition. In the next Section, an extended feature is created by combining LBP-TOP with Gaussian Derivatives (GD). The aim of this features is to improve the representation of micro-movements in LBP-TOP by using first and second order GD [155] to highlight facial features allowing the machine learning methods to better distinguish between them.

4.3 LBP-TOP with Gaussian Derivative Feature

This section describes the extended feature of LBP-TOP with GD and a method of differentiating between a micro-facial movement and a neutral expression. As a well established feature extraction method that focuses on textures [94], LBP-TOP

has been suited to highlighting temporal features effectively. To further improve and enhance the feature by getting a complete account of an image's grey-value-invariant structure [156], GD are applied to the LBP-TOP image. Further evidence to the effectiveness of GD is its use in the stages of SIFT [157], a very well established and popular way of detecting and describing local features in images.

Normalisation is described by automatically using the centre point of the two eyes and affine transformation to rotate each face from CASME II [17], and then cropping each image to just the face itself. Finally, LBP-TOP [94] and GD features are obtained and classified into either a micro-facial movement or neutral expression using RF and SVM.

4.3.1 Normalisation

Normalisation is applied to all sequences so that all the faces are in the same position based on a constant reference point, in this case, the midpoint between the eyes. Once the midpoint has been obtained, affine transformation is used to rotate the face so that all faces line up horizontally based on this point. The face of the sequences then needs to be cropped to remove the unnecessary background in each image.

To calculate the midpoint of the eyes, first the centre of both eyes are obtained automatically by using a Viola-Jones Haar cascade detector [58] to detect both the left and right eyes separately. Closed eye Haar detectors are available, however as the dataset does not include closed eyes, this has not been implemented. This creates a bounding box around both eyes which the centre point of an eye can then be extracted

$$\breve{C}_x = \frac{W}{2} + x \tag{4.1}$$

$$\breve{C}_y = \frac{H}{2} + y \tag{4.2}$$

where $\breve{C}$ is the centre of the eye, W is the width of the bounding box, and H is the height and x and y are the pixel locations of the top-left corner of the bounding box for the eye. Once the centre points are found for both the left and the right eye, this paper computes the midpoint of the eyes

$$\breve{M}_x = \frac{\breve{L}c_x + \breve{R}c_x}{2} \tag{4.3}$$

$$\breve{M}_y = \frac{\breve{L}c_y + \breve{R}c_y}{2} \tag{4.4}$$

where $\breve{M}$ is the midpoint between the eyes and $\breve{L}c$ and $\breve{R}c$ are the centres of the left and right eye respectively. Using the calculated points, affine transformation can be applied to all images. First the distance between the eyes is found by

$$\breve{D}_x = |\breve{R}c_x - \breve{L}c_x| \tag{4.5}$$

$$\breve{D}_y = |\breve{R}c_y - \breve{L}c_y| \tag{4.6}$$

where $\breve{D}$ is the distance between the eyes in the x and y coordinates. then the angle between the eyes is calculated by

$$\theta = \frac{\arctan(\breve{D}_x, \breve{D}_y)180}{\pi} \tag{4.7}$$

where θ is the angle between the eyes. Using the extracted points, affine transform is used to align the eyes horizontally, ready to be processed.

4.3.2 Processing Images

Feature extraction begins by grey scaling each image sequence and dividing each image into 9×8 non-overlapping blocks, as proposed by Zhao et al. [94] as their best performing block size (see Fig. 4.7). This sequence then has a GD operator applied with σ (the Standard Deviation (SD)) being changed from 1-7 in each iteration once the whole database has been processed.

To extract features such as blobs and corners from the face images, the first and second order derivatives [155] of the Gaussian function are calculated. The first order GD is defined as

$$G_x(x, y; \sigma) = \frac{\partial G(x, y; \sigma)}{\partial x} = -\frac{x}{\sigma^2} G(x, y; \sigma) \tag{4.8}$$

$$G_y(x, y; \sigma) = \frac{\partial G(x, y; \sigma)}{\partial y} = -\frac{y}{\sigma^2} G(x, y; \sigma) \tag{4.9}$$

where σ is the scaling element of the GD. The second order GD is defined as

$$G_{xx}(x, y; \sigma) = \left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right) G(x, y; \sigma) \tag{4.10}$$

FIGURE 4.7: Images are split into 9×8 blocks so each can be processed separately and obtain local features that are concatenated to form the overall global feature description.

$$G_{yy}(x, y; \sigma) = \left(\frac{y^2}{\sigma^4} - \frac{1}{\sigma^2}\right) G(x, y; \sigma) \tag{4.11}$$

$$G_{xy}(x, y; \sigma) = \frac{xy}{\sigma^4} G(x, y; \sigma) \tag{4.12}$$

the first and second order derivative features are then summed together to get the final GD feature and form a stronger feature representation of blobs, corners and other important features.

The chosen parameters used for micro-movement detection were recommended from the respective literature where first created - $LBPTOP_{8,8,8,3,3,4}$.

4.4 Experimental Results and Discussion

To test this method's performance, combinations of image planes are used with temporal and spatial mixes. The testing data is set up to 50%, therefore if 30%

is training the remaining 70% is used for testing. No data within the training set is used for testing to ensure all testing data is unseen. Each plane is tested using 10-fold cross-validation. Other literature [17, 64, 94] use leave-one-subject-out evaluation with data. This method uses more or equal testing than training to describe the robustness of this method compared to others in the literature. The dataset being used is the CASME II recorded at 200 fps with 35 Chinese participants with a mean age of 22.03 years.

For both RF and SVM the σ value for GD goes from $1-7$. In RF the accuracy increases until the 5th value, where it peaks and begins to decrease, indicating that when $\sigma = 5$ the accuracy is at its highest. In SVM, the accuracy decreases as the σ value increases.

Table 4.3 shows the results from the SVM experiment and Table 4.4 shows results from the RF experiment. SVM and RF results vary considerably with the highest accuracy for SVM was 54.3% with training set to 10%. The accuracy gradually decreased as training increased. As the data is high-dimensional and values lie close together, SVM struggles to separate the data beyond chance. As the training data increases for the SVM, the training model becomes overly complex and overfitting occurs.

As RF uses a bootstrap method it is able to generate many classifiers (ensemble learning) and aggregate results to handle the data more appropriately, only ever choosing random samples and ignoring irrelevant descriptors. This gave the highest accuracy of **92.6%** in the XTYT plane with a SD of 1.78. The main reason for LBP-TOP with GD performing better than LBP-TOP alone is how GD describe the grey-value structure of images in a scale and rotation invariant way [156].

By removing the spatial information and just using the temporal planes, classification results for RF are higher. In SVM the results did not vary considerably across planes, and the highest result was for all planes (54.3%, SD: 0.56) and the lowest being the XY plane alone (34.73%, SD: 2.6).

The highest results were found to be when the σ value was set to 5. Fig. 4.8 shows the gradual increase in accuracy as training is increased in all planes with a temporal element. A decrease was shown in just the XY plane, supporting that as more training is introduced, the XY plane acts as noise to any movement. This

TABLE 4.3: All results using the SVM classifier. Each plane used the the combination of LBP-TOP and GD features. The training percentage is displayed for each plane from 10% to 50%.

Plane	σ	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	1	51.20	47.10	43.30	39.90	36.20
XY	1	51.10	46.90	43.10	39.10	35.10
XTYT	1	51.90	47.90	44.20	40.00	36.80
YT	1	51.00	46.90	43.10	39.40	35.80
All Planes	1	52.50	48.70	44.80	40.90	37.50
XT	2	52.40	48.80	44.80	41.10	36.90
XY	2	52.10	48.10	44.30	40.20	36.40
XTYT	2	53.20	49.80	46.80	43.60	40.90
YT	2	52.50	48.80	45.10	41.30	37.70
All Planes	2	53.70	51.30	48.80	46.30	43.90
XT	3	52.30	48.50	44.60	41.00	37.20
XY	3	52.30	48.30	44.50	40.70	37.10
XTYT	3	53.20	50.20	47.50	44.70	41.70
YT	3	52.60	48.70	45.50	41.60	38.50
All Planes	3	54.20	52.00	50.10	48.30	46.20
XT	4	52.30	48.20	44.40	40.90	36.90
XY	4	52.40	48.30	44.60	40.70	37.10
XTYT	4	53.30	50.20	47.40	44.30	41.20
YT	4	52.40	48.70	45.30	41.90	38.50
All Planes	4	***54.30***	52.30	50.20	48.40	46.60
XT	5	49.82	46.33	42.47	39.56	36.21
XY	5	50.62	46.45	42.74	39.12	35.65
XTYT	5	51.51	47.36	44.03	40.18	37.00
YT	5	50.00	46.12	42.94	40.20	36.56
All Planes	5	52.28	48.26	44.32	40.62	36.40
XT	6	49.89	45.64	42.59	39.03	35.70
XY	6	50.47	46.04	42.47	38.66	35.02
XTYT	6	51.08	47.17	43.64	39.78	36.37
YT	6	49.84	45.86	42.38	38.95	36.11
All Planes	6	52.24	48.08	44.02	39.91	36.26
XT	7	49.62	45.40	41.87	38.27	34.93
XY	7	50.30	46.02	42.26	38.47	34.73
XTYT	7	50.81	46.79	43.36	39.43	36.13
YT	7	49.75	45.87	42.53	39.25	36.43
All Planes	7	52.14	48.01	43.91	39.83	36.09

TABLE 4.4: All results using the RF classifier. Each plane used the the combination of LBP-TOP and GD features. The training percentage is displayed for each plane from 10% to 50%. The results for RF are significantly higher than SVM with results starting to plateau and decrease when $\sigma = 6$.

Plane	σ	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	1	59.00	63.00	65.60	67.00	70.60
XY	1	51.20	49.30	48.10	46.30	44.50
XTYT	1	57.60	61.20	63.80	66.00	68.00
YT	1	56.60	59.00	60.50	62.70	65.30
All Planes	1	55.80	57.60	58.60	60.00	60.70
XT	2	66.80	73.70	77.20	80.70	82.30
XY	2	51.80	49.80	48.20	45.60	43.90
XTYT	2	66.10	72.20	75.90	79.30	81.60
YT	2	64.90	70.70	75.00	77.70	80.30
All Planes	2	61.30	66.40	69.50	71.20	74.00
XT	3	74.30	82.80	85.90	88.50	90.10
XY	3	52.90	50.80	49.40	48.30	46.20
XTYT	3	73.90	81.80	85.40	87.90	89.20
YT	3	72.30	80.20	84.30	86.50	88.00
All Planes	3	68.10	74.70	78.60	81.30	83.80
XT	4	79.40	86.80	89.20	91.30	92.40
XY	4	53.10	51.70	50.10	48.40	46.60
XTYT	4	78.50	86.10	88.50	90.90	91.70
YT	4	77.80	84.60	87.40	89.00	90.80
All Planes	4	70.60	78.10	81.70	84.80	86.70
XT	5	78.80	86.50	89.50	91.20	92.50
XY	5	53.30	51.50	49.80	47.10	45.40
XTYT	5	79.30	86.70	89.20	91.40	***92.60***
YT	5	78.30	85.70	88.70	90.60	92.20
All Planes	5	71.70	79.00	82.50	84.80	87.30
XT	6	78.60	85.70	88.70	90.80	91.80
XY	6	52.80	50.60	48.30	46.90	44.50
XTYT	6	78.30	86.10	88.70	90.90	92.00
YT	6	78.40	84.90	87.60	90.00	91.40
All Planes	6	70.30	77.30	80.40	84.30	86.40
XT	7	75.40	83.00	85.90	88.40	89.80
XY	7	52.70	50.20	48.30	45.40	42.80
XTYT	7	77.40	83.90	87.10	89.10	90.60
YT	7	77.60	83.90	87.20	88.80	90.20
All Planes	7	69.20	75.60	79.00	81.00	84.00

TABLE 4.5: All results using the SVM classifier when using only LBP-TOP features. The training percentage is displayed for each plane from 10% to 50%.

Plane	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	52.4	48.8	46.1	43	40.9
XY	51.9	48.1	43.9	40.1	36.7
XTYT	53.6	51.1	48.8	46.7	44.2
YT	53.1	50.6	47.9	45	43.1
All Planes	54.2	52.2	50.2	48.3	46.3

TABLE 4.6: All results using the RF classifier when using only LBP-TOP features. The training percentage is displayed for each plane from 10% to 50%.

Plane	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	60.4	64.3	66.6	69.4	71.4
XY	50.9	48.1	46	43.3	41.4
XTYT	58.8	61.7	63.3	65.5	67.8
YT	56.8	58.7	60.6	62.3	64
All Planes	56	57.8	58.5	59.6	59.9

can also be seen when all planes are used and the accuracy is pushed lower than just the temporal planes.

SVM and RF were also used to classify the image sequences using only LBP-TOP features. The results in Table 4.5 show that all of the planes perform no much better than chance, if not lower, with accuracy decreasing as the amount of training data is introduced. SVM appears to perform similar to results with GD, and separating the features is difficult.

Table 4.6 shows the results from RF using only LBP-TOP features. The accuracy for detecting movement increased significantly compared with SVM, however the highest result was lower than when combined with GD at **71.4%** when using 50% training and 50% testing data in the XT plane.

There has not yet been any literature from purely detecting micro-facial movement when comparing with neutral faces, or non-movements. Therefore, a benchmark for comparing our results could not be found. Most previous work focuses on detecting the movements and classifying into distinct emotional categories and therefore include automatic interpretation based on the FACS equivalent muscle movements (i.e. happy would be movement in AU12).

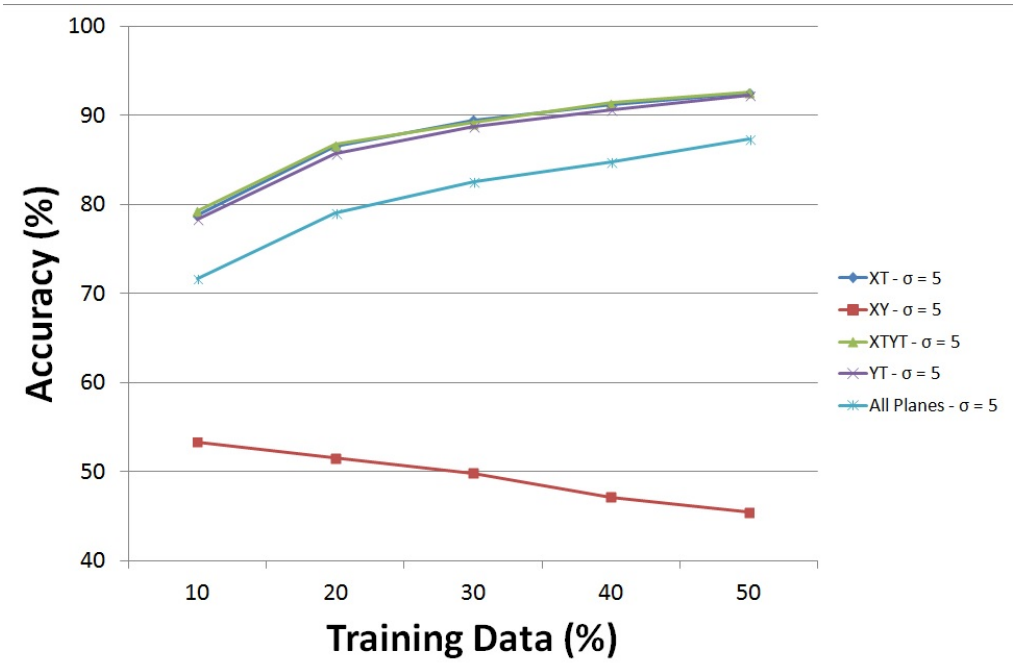

FIGURE 4.8: Using RF, the accuracy of all planes where $\sigma = 5$. Notice XY decreases as training increases due to the lack of temporal information.

After the preliminary tests (see Section 4.2), it was found that separating features into distinct micro-movement and non-movement classes was not a trivial task. Due to the subtle nature of the movements, the feature vary in small amounts between classes. Even though some results using RF performed quite well, the reality of too many false positives and inseparable data means machine learning is not suited to defining micro-facial movements.

A visualisation from Weka of the feature points when using the 3D HOG feature can be seen in Fig. 4.9. It shows a 3D plot of sample features using 3D HOG. The blue and red points show the micro-movement and non-micro-movement features respectively. The visualisation shows the lack of linear separation between features, with the points almost being in a random mix. However, a slight internal separation can be seen with more red points to the left and blue points to the right.

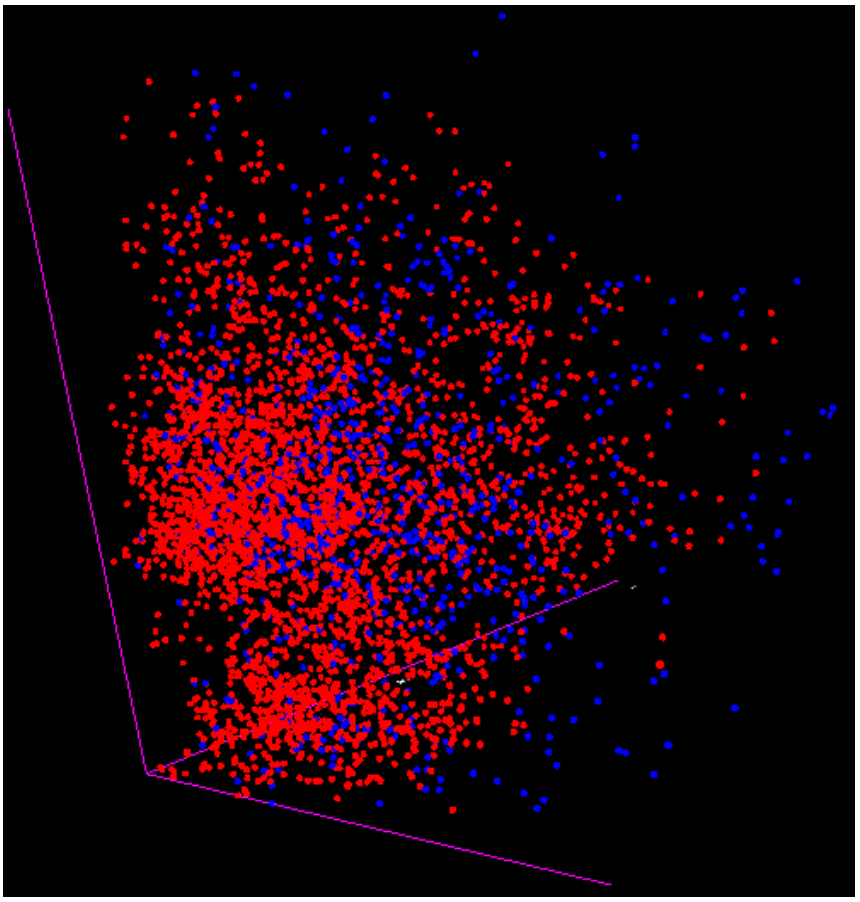

FIGURE 4.9: A 3D graph visualisation (generated in Weka [153]) of the feature points when using the 3D HOG feature. It is obvious that the data points have very little separation.

4.5 Summary

This Chapter proposes an extended feature, namely LBP-TOP with GD, to combine a well established spatio-temporal feature with first and second order derivatives. LBP-TOP represented micro-movements well in initial tests, so GD were incorporated to highlight facial feature for better machine learning classification. Results were promising, achieving the highest result of 92.6% using RF, however did not perform as well, even as training data was increased.

Early work on micro-expressions in computer vision involved machine learning and the attempt to recognise them into specific classes like normal facial expressions. Even though some results in this field have been reasonable, measure such as simple accuracy have been used, which do not take into account false positives or true negatives. From these findings, micro-facial expressions are very difficult to split into clear cut classes like positive and negative, if not impossible, with

the amount of variability with human emotions. An approach that objectifies the micro-expressions as movements s required, where the movements are treated as what they are: muscle movement of the face.

During investigations and creation of a novel feature, it was found that the amount of datasets that contained spontaneous micro-facial movements was lacking in number. To improve the research of subsequent Chapters and the field of micro-movement analysis, the SAMM dataset was created. This mainly expanded the demographic and resolution of any other micro-movement dataset available.

Chapter 5

Spontaneous Activity of Micro-Movements Dataset

This Chapter describes a new dataset named the Spontaneous Activity of Micro-Movements (SAMM). Currently, there are a limited number of micro-facial movement datasets, and the ones that do exist have their limitations. The new dataset will allow for more robust validation on real-world applications and contribute to the expanding field.

5.1 Introduction

Micro-movement datasets are currently very limited, as inducing these movements spontaneously is much more difficult than posed macro-facial expressions. Standardised micro-movement datasets are required to investigate the effectiveness of micro-movement analysis algorithms, including detection and recognition. Further, as more techniques are created, it is more difficult to compare algorithms without relevant standard datasets. The datasets currently available do not yet follow a clear standard, and so there are limitations when validating techniques.

This dataset is the first high resolution dataset, with inducement based on the 7 basic emotions [10] recorded at 200 fps. As part of the experimental design, each video stimuli was tailored to each participant, rather than getting self-reports after the experiment. This allowed for particular videos to be chosen and shown to participants for optimal inducement potential.

5.2 Experimental Protocol

The experiment comprised of 7 stimuli used to induce emotion in the participants who were told to suppress their emotions so that micro-facial movements might occur. To increase the chance of this happening, a prize of £50 was offered to the participant that could hide their emotion the best, therefore introducing a high-stakes situation [10, 41]. Each participant completed a questionnaire prior to the experiment so that the stimuli could be tailored to each individual to increase the chances of emotional arousal.

5.2.1 Emotion Inducement Procedure

Participants were first introduced to the experiment, and each were asked if they have read the participant information. A release agreement was signed and the participant was shown to their seat. The observer let the participant know that they could stop at any time, due to the potential for the stimuli to over-stimulate their emotions. Participants were also reminded that they were to suppress their true emotions and keep a neutral face with the aim of winning £50. The observer then moved to the observation room and the experiment began.

Each stimuli was shown and the participants were asked after every one if they were happy to continue, this ensured participants fully offset from any emotion they were feeling. Participants were only recorded when the stimulus was shown. After recording the observer returned to the experiment room and thanked the participant. A summary of the data collection procedure can be seen in Fig. 5.1.

Suppression of emotions is inherently a social act, and keeping the observer a short distance away in an observation room may seem opposite to this. However, due to the lab setting that participants are within, participants may not be fully relaxed if they are constantly aware of someone watching. The observer is kept out of sight to maximise the chances of natural suppression by making participants as comfortable as possible.

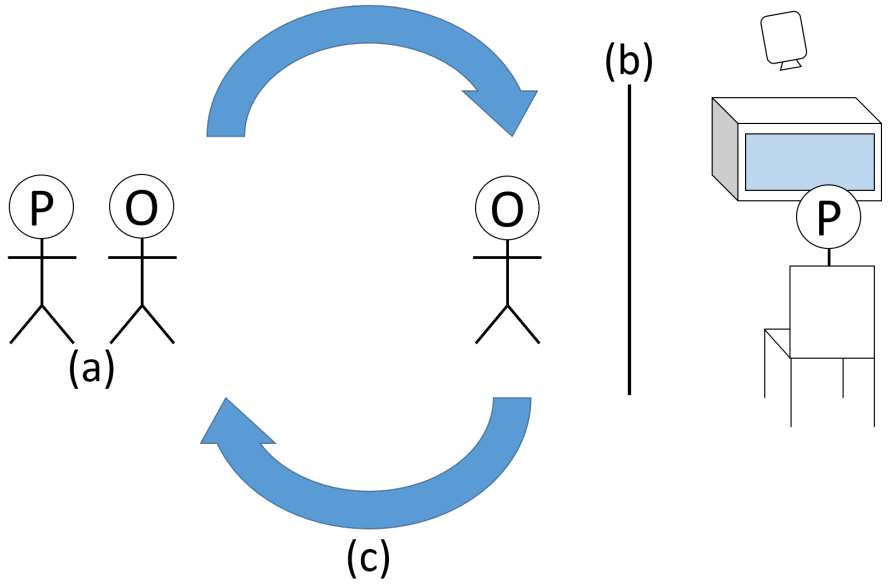

FIGURE 5.1: (a) The participant (**P**) is greeted by the observer (**O**) and gives an overview of the experiment. (b) The **P** completes the experiment by watching the 7 tailored stimuli while the **O** records the **P** using a high-speed camera. (c) The **P** is thanked and leaves the experiment. The next **P** is then invited into the room to begin the experiment again.

5.2.2 Ethics

A formal ethical clearing procedure took place to safeguard participant's when experiencing stimuli that provokes an emotional response. Every person was free to stop the experiment at any time and a full information sheet and release agreement on how the data would be used was issued. The study was approved by the Research Ethics Committee at Manchester Metropolitan University (approval SE121318A1). All participants gave signed informed consent to take part in data collection and for their images to be published.

5.2.3 Equipment and Experimental Set-Up

The experiment was setup in a room that helped keep interaction with the participant to minimum and allow them to become comfortable in the laboratory surroundings. The observer controlled the equipment in one room and the participant had emotional stimuli shown to them to induce particular emotions (see Fig. 5.2). A one-way mirror allows observation of the participants and an intercom

FIGURE 5.2: The left hand room shows the observation room and the right side shows where the participant completes the experiment.

system was used to communicate with participants, if necessary, without physically entering the room to keep interruption to a minimum. Participants watch all emotional stimuli on a 32 inch flat-screen television.

The experiment room contained all the equipment required for capturing the high-speed videos. To set up the environment, the camera and participant chair stayed in the same position for every person, however the lights required to be adjusted based on a person's height to ensure an even lighting on the face. The camera is connected to a system that is able to continuously capture high-speed video data for up to 20 minutes.

5.2.3.1 Camera

The camera used was a Basler Ace acA2000-340km, with a grey-scale sensor, set to record at 200 fps. The resolution was set to the highest possible: 2040×1088 pixels, and is currently the highest resolution available for this type of dataset.

FIGURE 5.3: The Basler Ace acA2000-340km high-speed camera used to capture videos in the emotional inducement experiment.

5.2.3.2 Lighting

Lighting can be problematic for high-speed cameras as many lighting systems use alternating current that refreshes regularly at a usual frequency of 50 Hz. Recording at 200 fps, the camera can pick-up the lights refreshing and this shows as flickering on the resulting images. To counter this, two lights that contained an array of Light Emitting Diodes (LEDs) was used with an illuminance of 1750 lux at 50cm. As the lights are plugged into the mains electricity, we had to ensure that direct current was still being used to avoid the flickering on the final images. Light diffusers were placed around the lights to soften and even out the light on participant's faces.

5.2.3.3 High-Speed Data Capture

The system used to capture the images uses a frame grabber and a RAID array of independently tested solid state drives to ensure no dropped frames occur. The software used was IO Industries Streams 7 that allows for recording and analysis of the data. As the software initially records to a proprietary format, the original can be used to export various formats as required.

5.2.4 Image Noise Considerations

When creating a dataset with digital videos, noise created from a variety of factors can occur. Principle sources of noise in digital imagery are from acquisition and/or transmission [158]. The performance of the acquisition device can be affected by a variety of factors during capture, including environmental conditions, light levels (low light conditions require high gain amplification), sensor temperature (higher temperatures imply more amplification noise). Images can also be corrupted during transmission due to interference in the channel due to electrical issues, computer interface problems or atmospheric conditions [159].

Acquisition issues are the main concern for the high-speed camera used in this experiment. Lighting is one of the most important factors, as high-speed cameras need more lighting due to the higher the frame rate and shutter speed. These issues were solved in Section 5.2.3.2. The Basler camera used is also very small (see Fig. 5.3), meaning that constant use increases the temperature of the camera considerably. To minimise the risk of acquisition problems, the camera was powered down when not in use between participants.

Most sources of noise is generally characterised as additive Gaussian [160], however Poisson noise can be found when video sequences are acquired under quantum-limited conditions. Although every care was taken to reduce the noise from the SAMM dataset, it is impossible to capture high-speed videos with absolutely no noise. Other methods of de-noising include treating the image sequence as a 3D volume, and applying various transforms to this volume in order to attenuate the noise [161].

5.2.5 Inducement Stimuli

The majority of the emotional inducement stimuli were video clips from the Internet. If a participant was fearful of heights, a first-person video of someone bungee jumping was shown. Further information on the tailored videos are discussed in the questionnaire section and a description of the video clips used is shown in Table 5.1 along with the emotion linked to the inducement.

The one stimulus that was different to a video clip was for surprise, where a presentation was used and shown as the last stimulus. The presentation appeared

TABLE 5.1: Tailored Stimuli Used to Induce Emotions

Video Stimuli Description	**Duration**	**Emotion Link**
Westboro baptist church	0'50"	Contempt
Lancing a boil	0'21"	Disgust
Snake attacks camera	0'17"	Fear
Angry dog barking through a fence	0'24"	Fear
Jumping spider	0'25"	Fear
Attacking crab	0'19"	Fear
Scary puppets	0'17"	Fear
Large spider chase	0'17"	Fear
First person bungee jump	0'16"	Fear
Moth flying around	0'21"	Fear
Racist woman	0'25"	Anger
A dog being kicked	0'26"	Anger
Bullying	0'35"	Anger
Unruly teenagers	0'28"	Anger
Movie death (Champ)	0'46"	Sadness
Bullying	0'35"	Sadness
A dog being kicked	0'26"	Sadness
Twin towers collapsing	0'49"	Sadness
Baby laughing	0'26"	Happiness
Flight of the Conchords song	0'22"	Happiness
Dog biting	0'14"	Happiness
Presentation with Participant's face	N/A	Surprise

to be boring slides that used a lot of text, however within the slides was an image of the participant. This enabled an unexpected event, without the risk of startling the participant.

5.2.6 Questionnaire

Some datasets [15–17] assign an emotion label to videos based on self-reports completed by participants at the end of the experiment. Therefore participants wait until a complete set of stimuli have been experienced and then get asked what emotion they felt during each stimulus.

TABLE 5.2: Questionnaire Participants Filled in Before the Experiment

Question No.	Question
1	What are your current fears and phobias?
2	Highlight the principles or moral standpoints that you hold
3	How is your view of another person / group affected when they have opposite moral standpoints to your own?
4	Outline the beliefs and values that you hold
5	Describe what makes you angry
6	Describe what makes you sad
7	Describe what makes you disgusted
8	Describe what makes you happy

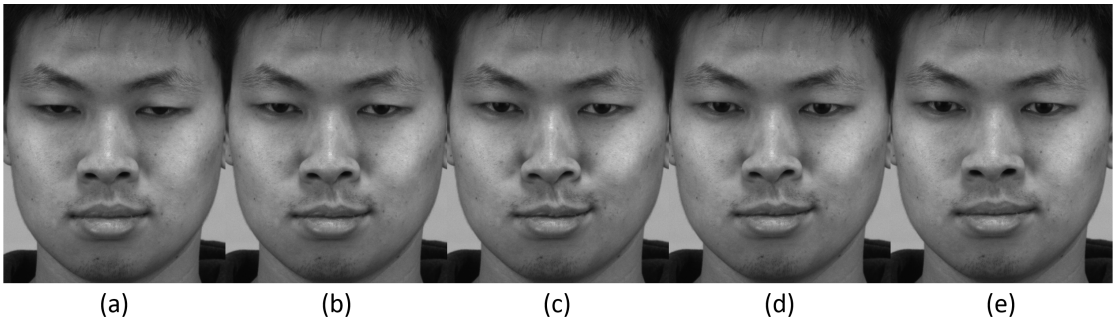

FIGURE 5.4: An example of a coded micro-movement. The movement shows AU 13 and AU 39, which is the sharp lid puller and nostrils compressing. Image (a) is the onset frame, (c) is the apex where the mouth curls upwards sharply and the nostrils move inwards. Finally, (e) is the offset frame.

To minimise the risk of participants forgetting or making up their report, each person was asked to fill in a questionnaire before turning up to the experiment so that each emotional stimuli video could be tailored to what each person found to induce emotion in themselves. All of the questions can be found in Table 5.2.

5.2.7 FACS Coding - Ground Truth

To obtain ground truth every movement was coded using FACS and inconsistencies between coders eliminated by mutual cross-checks to establish a consensus. This is especially relevant for micro-movements as coding is done objectively based on the muscle movements, and does not try to interpret an emotion [39].

The FACS coding was completed by three certified coders, who were not otherwise involved in the study, to ensure inter-coder reliability. Coding was performed after the videos have been recorded in accordance with usual FACS coding procedure. At no point did the coders know the context of the videos they were coding, which means no coder was aware of the stimuli used to induce emotion in participants.

The inter-coder reliability of the FACS codes within the dataset is 0.82, and was calculated by using a slightly modified version of the inter-reliability formula found in the FACS Investigator's Guide [40]. It is defined as

$$Re = \frac{3(AU(C_1, C_2, C_3))}{All\_AU} \tag{5.1}$$

where Re is the reliability score, $AU(C_1, C_2, C_3)$ is the number of AUs where all coders agreed on the same AU and All_AU is the total number of AUs scored by all three coders. In contrast, other FACS coded datasets usually have two FACS coders to code each video, however to increase reliability and ensure accurate ground truth, three coders were used in this dataset.

5.3 Dataset Analysis

The SAMM dataset contains micro-movements captured at 200 fps. All macro-movements were also coded as to not disregard potential useful movements that may be used at a later date. Every movement was coded, including the macro-movement, with an onset, apex and offset frame to note the duration. Each micro-movement is coded with micro-movements lasting 100 frames or less, translating to 500 ms at 200 fps. Any movements that were coded to be longer than 100 frames in duration would be classed as a macro-facial expression. An example micro-movement that has been FACS coded from the SAMM dataset can be seen in Fig. 5.4.

Unlike most other datasets, every FACS AU was coded regardless of their relation to emotions. This includes head and eye movement codes. By FACS coding the data comprehensively, the dataset can be used for much wider purposes than exclusively for micro-movement analysis.

TABLE 5.3: Participant Age Distribution

Age (Yrs)	Male	Female	#Participants
19 - 29	9	6	15
30 - 39	5	3	8
40 - 49	1	4	5
50 - 59	1	3	4

TABLE 5.4: Ethnicity/Race of Participants

Ethnicity/Race	Total
African	1
Afro-Caribbean	1
Arab	2
Black British	1
White British	17
White British / Arab	1
Chinese	3
Indian	1
Malay	2
Nepalese	1
Pakistani	1
Spanish	1

5.3.1 Demographic Breakdown

To obtain a wide variety of emotional responses, the dataset was required to be as diverse as possible. A total of 32 participants were recruited for the experiment from within the university with a mean age of 33.24 years (SD: 11.32, ages between 19 and 57). An even gender split was achieved, with 16 male and female participants. Table 5.3 shows the age distribution and gender groupings of participants, and Table 5.4 summarises the ethnicities and the respective counts in each group.

Statistics for micro-movement AUs were calculated for two categories of duration:

- Up to 100 frames (i.e. half a second).

TABLE 5.5: The occurrence frequency for both duration groups has been calculated for the main upper and lower face AUs.

Upper Face			**Lower Face**		
AU	**No. Of Occurrences**		**AU**	**No. Of Occurrences**	
	Up to 100 Frames	**101 to 166 Frames**		**Up to 100 Frames**	**101 to 166 Frames**
1	6	5	9	5	1
2	16	7	10	5	3
4	23	14	12	29	13
5	9	8	14	11	7
6	5	0	15	4	1
7	45	14	17	7	6
Other	9	9	20	7	2
			23	1	3
			Other	40	23
Total	113	57	**Total**	109	59

- From 101 to 166 frames (i.e. two-thirds a second).

Using up to 100 frames allows for comparison against CASME II, which labelled their data to this length. Additional statistics for the second group can be used for when the duration of the movement is defined slightly higher than usual. Table 5.5 outlines the frequency occurrences for well known AUs in these groups. There was a total of 222 AUs in the group of up to 100 frames and 116 in the group up to 166 frames. A large portion of the overall AUs coded in the dataset turned out to be in the micro-movement category. The percentage of all 338 micro-movements in both duration categories (shown in Table 5.5) was 45.3%, and only up to 100 frames was 29.7%.

5.3.2 Statistical Analysis

A chi-square (χ^2) test was conducted using all observed facial movements to test the significance of the different AUs invoked by the emotional context. Certain FACS AUs are used within Emotion FACS (EMFACS) [40] to only describe critical AUs related to emotion. Table 5.6 shows occurrences of these key reliable muscle movements during specific stimuli. Non-reliable muscle movements have been included for statistical analysis to show AUs that are not classed as reliable, but occurred frequently across participants.

The reliable muscles for Contempt did not occur during any stimuli, and so this group has not been included. The Surprise reliable group has been included,

TABLE 5.6: Frequency occurrences of reliable AUs pooled together to form AU groups. The values correspond to how many times a group occurred when a participant was shown a particular stimulus category. Also shown is the non-reliable movements that do not relate to emotional context, but occur frequently.

Stimuli Category	**Reliable AU Groups**							
	Disgust	**Fear**	**Anger**	**Sadness**	**Happiness**	**Surprise**	**Non-Reliable Movements**	**Total**
Contempt	5	1	9	4	15	1	54	**89**
Disgust	12	2	12	5	17	0	43	**91**
Fear	7	4	5	10	31	1	41	**99**
Anger	5	0	8	1	10	0	67	**91**
Sadness	1	4	3	10	10	5	53	**86**
Happiness	10	4	5	14	69	2	39	**143**
Surprise	8	3	4	15	37	4	75	**146**
	48	**18**	**46**	**59**	**189**	**13**	**372**	**745**

but has too few results to allow for reliable statistical comparison and has been omitted from calculations. The data for each individual AU has been pooled into categories for the χ^2 test to be acceptable and the significance level was set to $\alpha = 0.05$.

An investigation is undertaken to find if the internal values in Table 5.6 can be predicted from the marginal totals. The statistical χ^2 equation can be defined as

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \tag{5.2}$$

where the actual *expected* and *observed* frequency of AUs is analysed. For only reliable AU groups, the $\chi^2 = 82.28$, $p = 9.25 \times 10^{-7}$ and Degrees of Freedom (df) = 30. When the non-reliable movements are included the $\chi^2 = 136.4$, $p = 1.34 \times 10^{-13}$ and df = 36.

From this analysis, the hypothesis with no association between facial movements and stimuli can be rejected as there is statistical significance between the two. Further to this conclusion, from the observed values the pooled Happiness reliable AUs and stimuli have the highest frequency and show a correlation between the movement and emotional context. In other groups this is less apparent, however unlike in similar experiments performed by Mammucari et al. [162] the experimental protocols required participants to suppress their true emotion, therefore making it less likely, if the experiment was a success, for participants to show all reliable muscles. For example, some participants showed a single AU rather than combinations of AUs, and the masking of reliable AUs with other movements is a side-effect of asking participants to suppress.

5.4 Dataset Validation Tests

5.4.1 Results Methodology

Four spatio-temporal methods were used to perform initial tests on the dataset and provide results to compare previous methods that use machine learning classification on SAMM. The first two methods are LBP-TOP based and the others are HOG based. One of the HOG-based sliding window descriptors by Felzenszwalb et al. [163] and is also used, and is named Deformable Part Models (DPM). They use star models defined by HOG as a coarse root filter that covers the entire object and then creates higher resolution part filters to describe local regions of an object (in this case the face). The parts, when described together, form an object only when they are meaningful in their geometrical constraints in the spatial domain originally, and extended into the temporal domain for this method.

Normalisation is applied to all sequences so that all the faces are in the same position by using affine transformation. The points used for alignment are obtained using the Face++ automatic facial point detector [164]. The face of the sequences then needs to be cropped to remove the unnecessary background in each image.

Feature extraction begins by grey-scaling each image sequence and dividing each image into blocks. To test different regions, images were divided into 4×4 and 5×5 regions with 16 and 25 video blocks for each movement respectively. Using 5×5 blocks allows for comparison with the CASME II procedure [17], and the other 4×4 blocks tests different local regions. Each video block then had the temporal descriptor applied as outlined in the previous section. Different blocking configurations may suit the aligned images better and changing the sizes allowed for testing this hypothesis.

Finally, the image sequences are classified using RF [26] with the default parameters in the machine learning tool, Weka [153]. Binary classification is used with the two classes being movement and non-movement. Each video block was assigned a ground truth label, from the FACS coding, and 10-fold cross validation was performed. This involved splitting the micro-movement and non-movement data into 10 parts (folds) and holding out each part testing once, and used for training the remaining nine times. Cross validation is preferred over than repeatedly holding out the same data as it reduces the variance of the estimate.

As all the features, such as HOG, had been split into bins, the normalised micro-movement and non-movement blocks were concatenated into columns corresponding to each bin. This loses temporal information but allows for input into Weka. The overall classification accuracy, the F-measure and MCC (see Section 3.6) for the movement class was then calculated.

5.4.2 Results

Preliminary results were good when testing this new dataset on existing temporal descriptors. Using binary classification, LBP-TOP with a radius of 3, 3, 3 for XY, XT and YT planes respectively and using a 4×4 block configuration produces the best result of **0.67** when using the F-measure statistic.

For binary classification, the MCC introduced by Matthews et al. [165] takes into account the true and false positives (TP and FP) and true and false negatives (TN and FN) to obtain a balanced coefficient measure between -1 and 1, where 1 is perfect classification, 0 is random chance and -1 is total disagreement.

Table 5.7 details the performance of the two spatio-temporal descriptors based on LBP. The results are slightly reduced when using LBP-TOP with GD compared with LBP-TOP. The reason for this is due to a change in datasets, which is likely to cause some minor fluctuations when compared with the performance of LBP-TOP with GD in Section 4.4. Table 5.8 shows the performances of 3D HOG and DPM, which are both based on Histogram of Oriented Gradients (HOG) descriptors.

In all cases, the descriptors are attempting to generalise the micro-movements and non-movement blocks across all instances and performs well for this difficult task. Further discussion on the generalisation of micro-movements will be within the next section.

5.5 Comparison with Current Datasets

Current datasets containing the first occurrence of micro-movements were first described in Section 2.8. As the SAMM dataset is the most recently created, it would be useful to compare the strengths and weaknesses against the datasets

TABLE 5.7: Results calculated using 10-fold cross validation. Spatio-temporal methods based on LBP are described using different LBP radii and different block splitting sizes. The final two results only use the XT plane and the best results are highlighted in bold.

		Block Configuration					
		4×4			5×5		
Descriptor	**Radii**	Accuracy (%)	F-Measure	MCC	Accuracy (%)	F-Measure	MCC
LBP-TOP	3,3,3	**82.70**	**0.67**	**0.56**	91.32	0.53	0.52
LBP-TOP	1,1,4	82.31	0.66	0.55	**91.52**	**0.53**	**0.53**
LBP-TOP & GD	3,3,3	82.07	0.65	0.54	90.80	0.47	0.47
LBP-TOP & GD	1,1,4	82.35	0.66	0.55	91.04	0.50	0.49
LBP-TOP & GD (XT Plane)	3,3,3	80.18	0.61	0.49	89.16	0.32	0.34
LBP-TOP & GD (XT Plane)	1,1,4	80.46	0.62	0.50	88.68	0.30	0.30

TABLE 5.8: Results calculated using 10-fold cross validation. Spatio-temporal methods for HOG and DPM are described using different block splitting sizes.

	Block Configuration					
	4×4			5×5		
Descriptor	Accuracy (%)	F-Measure	MCC	Accuracy (%)	F-Measure	MCC
3D HOG	79.05	0.59	0.46	90.36	0.42	0.43
DPM	78.85	0.56	0.45	90.14	0.41	0.42

TABLE 5.9: Summary of Publicly Available Datasets Containing Micro-Facial Movements

	Polikovsky et al. [5]	SMIC [15]	USF-HD [6]	CASME [16]	CASME II [17]	SAMM
Micro-Movements	42	164	100	195	247	159
Participants	10	16	N/A	35	35	32
Resolution	640×480	640×480	720×1280	640×480/720×1280	640×480	2040×1088
Facial Resolution	N/A	190×230	N/A	150×190	280×340	400×400
FPS	200	100	29.7	60	200	200
Spontaneous/Posed	Posed	Spontaneous	Posed	Spontaneous	Spontaneous	Spontaneous
FACS Coded	No	No	No	Yes	Yes	Yes
Emotion Classes	6	3	6	7	5	7
Mean Age (SD)	N/A	26.7 (N/A)	N/A	22.03 (SD = 1.60)	22.03 (SD = 1.60)	33.24 (SD = 11.32)
Ethnicities	3	3	N/A	1	1	12

described in the literature. A summary of all the current micro-movement datasets can be found in Table 5.9.

5.5.1 Polikovsky Dataset

The Polikovsky dataset, named after the author of the originating paper [5], was one of the first micro-movement datasets to be created. Unfortunately, this dataset is not publicly available, uses posed micro-expressions and only used 10 students

as participants. The SAMM dataset has a much higher resolution and has spontaneous micro-movements that are FACS coded.

One similarity is that it is one of a limited number of these datasets that recorded their participants at 200 fps. Further, a lot of consideration was taken to ensure the lighting was optimal so shadows were reduced. It also positions the camera so that the face is the main focus. A limitation of the SAMM dataset is that the face was not as close to the camera as hoped due to the lighting on the television being captured as flickering light.

5.5.2 USF-HD

Similarly to the previous dataset, the USF-HD [6] was also posed micro-expressions and is not available for public use. Also, it was recording at 29.7 fps, considerably lower than the 200 fps of the SAMM dataset. Such a low frame rate may fail to capture all the details of micro-movements, especially considering the speed could be around 1/5 of a second [18].

5.5.3 YorkDDT

This dataset is unique in this comparison, as the micro-expression data was extracted by Pfister et al. [64] from a psychological study named the YorkDDT [116]. This is in contrast to the other datasets that were created specifically for micro-movement research. Only 18 micro-facial expressions were extracted and no detail into the AUs or participant demographic. Even though the data was spontaneous, compared with SAMM the YorkDDT is extremely simple with little experimental protocols to study. It also has not been released for research purposes.

5.5.4 SMIC

The SMIC dataset [15] is a relatively recent dataset containing 164 spontaneous micro-facial movements. The SAMM dataset contains 159 micro-movements (up to 100 frames), giving an advantage over the SMIC. However, the frame rates of these videos are at 100 fps, meaning SAMM, recorded at 200 fps, provides more temporal information for analysing micro-movements.

The demographic is the SMIC dataset is reasonable, but not as widespread and diverse as SAMM. All movements were annotated with positive, negative and surprise classes instead of FACS coding. This makes the data much harder to use compared to the comprehensive coding provided in SAMM.

As this dataset uses high-speed videos, it is prone to image noise. Usually this can be solved with at least correct lighting procedures, however the SMIC contains flickering in many videos. One advantage over SAMM is that each participant was asked to report on how they felt after each stimuli they saw, whereas SAMM contains independent video ratings rather than from participants.

5.5.5 CASME and CASME II

The CASME [16] and CASME II [17] datasets contain the largest amount of micro-facial movements of all the datasets, with 195 and 247 micro-movements respectively. The main contributory factor in the larger amount of micro-movements compared to SAMM is that the CASME and CASME II datasets use a large selection video stimuli for every participant. Also, these videos are much longer (many over 1 minute) than the ones used in SAMM (all under 1 minute).

The CASME dataset was only recorded at 60 fps, so cannot be compared to SAMM and CASME II, which was recorded at 200 fps. Both dataset also used mostly students with a mean age of 22.03 years (SD = 1.60). SAMM has an advantage here in representing a population better with a mean age of 33.24 years (SD = 11.32, ages between 19 and 57). SAMM also improves on using multiple ethnicities compared to only using Chinese like CASME II.

As the CASME II is the most recent dataset released before SAMM, and is the most similar in terms of FACS coding and experimental protocols, it will be the primary dataset used in comparison for this work.

5.6 Summary

As a solution and contribution to the lack of spontaneously induced micro-facial movements, a new dataset is created. The contributions of this dataset include a wider demographic than other datasets to better represent a population. Ages

range from 19 - 57, there are 12 different ethnicities in total and the resolution is the highest of any other currently available dataset for micro-movements. Also, the SAMM dataset contains all the raw recordings from the experiment, which means sequences that do not have AUs assigned can be used as baseline sequences. No other dataset provides baselines as a main contribution. To summarise, SAMM contains largest amount of different ethnicities, resolution and age distribution of any other dataset of a similar kind currently available publicly.

Initial validation results were performed on the dataset to test it against descriptors previously used and to see how well micro-movements could be classified. Further tests are required, but current results show that the SAMM dataset can be used as a new benchmark dataset in micro-movement detection.

The exact experiment protocols have been defined to allow for easy reproducibility, potentially creating a standard for micro-movement inducement experiments. This new dataset can now be used in future experiments alongside others, and the addition of SAMM will help advance the research on micro-movement analysis.

Chapter 6

Micro-Movement Detection: Feature Difference Approach

This Chapter proposes a novel method of detecting micro-movements with temporal feature difference approach. This is achieved by calculating the difference between frames of a micro-movement and then using an individualised baseline feature to find the video frame where a micro-movement is found. The new baseline threshold is also developed into an Adaptive Baseline Threshold (ABT).

6.1 Introduction

Recent work into micro-movement detection has begun to use feature difference methods, that remove the need for machine learning and use feature descriptors, such as LBP, to find the feature changes between video frames. When feature difference changes occur, the temporal phases of the micro-movement are plotted graphically as peaks. If a peak crosses a certain threshold, then it is classed as a micro-movement, and a non-micro-movement otherwise.

Moilanen et al. [109] were the first to propose a feature difference method using LBP and a defined threshold based on the minimum and maximum values of the movement feature. Unfortunately, by doing this the calculated threshold would always a movement, even if it did not exist. The proposed method [166] solves this issue by taking into account the baseline of the participants of the

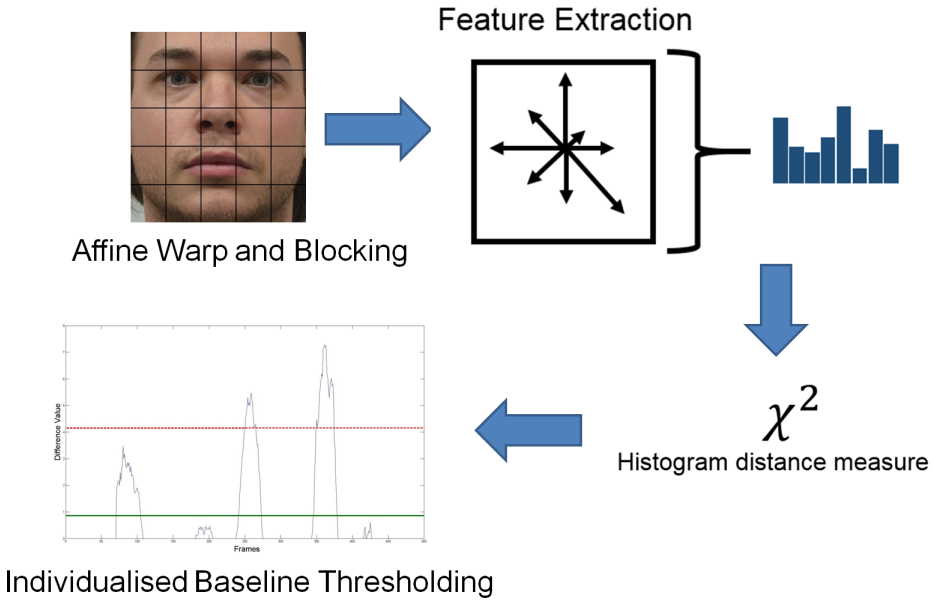

FIGURE 6.1: System summary of the feature difference approach and the proposed individualised baseline thresholds.

datasets (SAMM and CASME II) to determine what the threshold value should be. This follows a more psychological approach by using a person's baseline to detect changes [10], leading towards an objective system.

The baseline approach is also further expanded using an Adaptive Baseline Threshold (ABT) that finds a threshold value by taking both the baseline feature and movement feature of each individual participant. The threshold then adapts based the movement current being analysed. A summary of the overall system can be seen in Fig. 6.1.

6.2 Feature Difference

6.2.1 Preprocessing

Firstly, all sequences are normalised so that all the faces are in the same position by using affine transformation. The technique of aligning the faces was the same as in Section 4.3.1.

The face is divided into blocks to preserve the local texture and global shape of the face [148]. This method has worked well for LBP with facial analysis tasks, and so it will be applied to HOG features to analyse its effectiveness. HOG originally

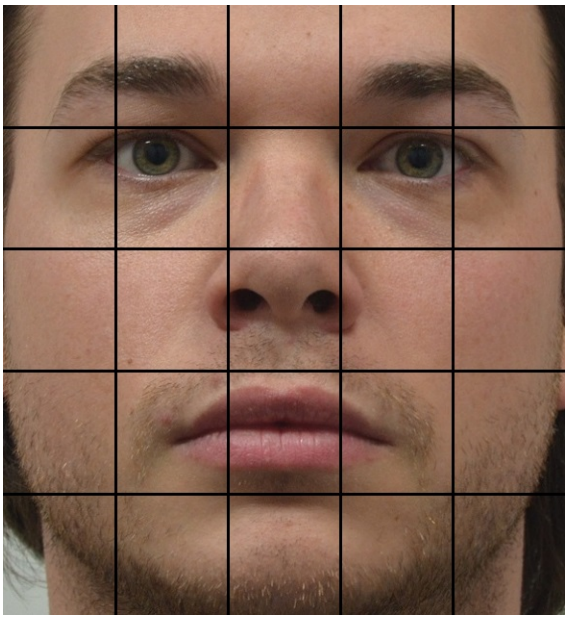

FIGURE 6.2: An example of a face divided into 5×5 blocks.

performs some sliding window, block-based normalisation to obtain a final feature, however we leave this out to keep a standard across descriptors.

6.2.2 Difference Analysis

With all videos captured, some form of noise is certain due to the way images are captured digitally, whether this be through lighting, temperature or equipment malfunction. High-speed video is particularly susceptible due to capturing lots of images in a short space of time. To counter this, de-noising is applied using the sparse signal processing method [139] of collaborative filtering.

Spatial HOG features are extracted from each frame of each block of the video using Piotr Dollár's Matlab Toolbox [142]. This is done similarly to Dalal and Triggs [11] in that the pixel orientation and magnitude values are calculated and then binned into a histogram based on the chosen orientations. The original gradient computation was used where Gaussian $\sigma = 0$ (no smoothing) and the derivative mask used is a simple 1D $\begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$ mask.

Originally the amount of orientation bins chosen in [11] are 9, but to model the movements of facial muscles in the spatial domain our method uses 8 bins. Also, unlike the original paper, the proposed method uses 2π for signed gradient

orientation binning so as to model the different directions the gradients can take and its relevance to facial expressions.

The chi-square (χ^2) distance [109] is then applied to obtain a feature vector of the de-noised HOG features of the SAMM dataset, and can be defined as

$$\chi^2(P,Q) = \sum_{b}^{B} \frac{(P_b - Q_b)^2}{(P_b + Q_b)} \tag{6.1}$$

where b is the b-th bin in the P and Q histograms that have an equal number of bins for a total amount of bins B.

6.3 Individualised Baseline Analysis

To address the limitations of the threshold calculated in [109] an individualised baseline threshold obtained from the neutral videos of participants is proposed. To provide a simple way of referring to the two types of threshold being discussed, the following conventions will be used. *T1* will refer to the threshold proposed by Moilanen et al. [109] and *T2* will refer to the individualised baseline threshold proposed by this research.

The *T2* value is computed by taking a neutral video sequence for the participants and using the χ^2 distance to get an initial feature for the baseline sequence. The maximum value of this baseline feature is then used as the *T2* value.

An example of the features extracted from a baseline video sequence and a movement video sequence (from the same participant) is illustrated in Fig. 6.3. The top blue line shows the histogram distances produced from the movement features and the bottom red line is the non-movement sequence. As can be seen, there are large deviations in histogram distances between the two temporal features.

6.4 Adaptive Baseline Threshold

The original baseline threshold calculation described in Section 6.3 took the maximum value of the baseline feature to set as the threshold value. This calculation was very simple, but quite effective in determining what was a micro-movement.

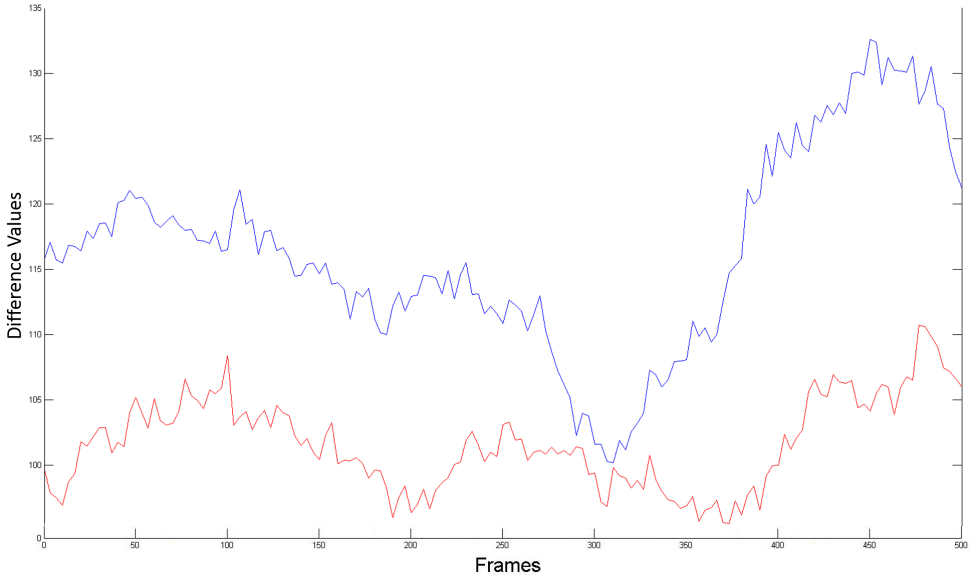

FIGURE 6.3: The raw feature difference of a micro-movement and baseline corresponding to that participant.

However, using the maximum value always set the threshold at the highest possible point. In doing this, there was many occasions when some very subtle movements were missed.

Instead of only using the maximum baseline feature value, an Adaptive Baseline Threshold (ABT) is proposed that takes into account the mean of both the movement and baseline feature vector. It can adapt the threshold level based on a balance between what is happening on the participant's face and what their baseline expression level is. The motivation behind an ABT comes from empirical experiments and finding a balance between the individual movement feature and baseline features that provide a fair threshold for determining a micro-movement has occurred. It also ensures the problems observed in Moilanen et al. [109], where the threshold will always detect a movement peak, do not occur.

If the maximum value of the baseline feature is greater than the mean of the current movement, then the threshold is set to this maximum. Otherwise, a balanced average is calculated between the mean value of the baseline and movement feature. The ABT can be calculated by

$$ABT = \begin{cases} max(\beta), & \text{if } max(\beta) > \bar{\epsilon} \\ \frac{\bar{\epsilon}+\bar{\beta}}{2}, & otherwise \end{cases} \tag{6.2}$$

where ABT is the calculated adaptive threshold, β is the baseline feature vector and $\bar{\beta}$ is its mean. The movement feature vector and its mean is denoted by ϵ and $\bar{\epsilon}$ respectively.

6.5 Experimental Results

The results shown in Table 6.1 outlines statistics on detection rates using the in-house dataset on the micro-movement detection using *T1* and the proposed method. Using *T2* calculated using the baseline of a participant's neutral facial expression, the results indicate a higher performance.

The proposed method is able to spot a large number of micro-movements when using a higher temporal resolution of 200 fps. Further it was concluded that the spatial HOG descriptor outperforms when compared using LBP. Further enhancements to the ABT could be to employ metaheuristic search algorithms to automatically adjust the threshold setting guided by a fitness function.

6.5.1 Measures of Performance

For these experiments, a reminder of the performance measure equations used are outlined. The *Precision* measure of exactness determines a fraction of relevant responses from results. It is defined as

$$Precision = \frac{TP}{TP + FP}. \tag{6.3}$$

Recall calculates the fraction of the results that are relevant to the experiment and that are successfully retrieved. It is commonly used with recall to form an understanding of the relevance of the results returned from experimental classification. It is defined as

$$Recall = \frac{TP}{TP + FN} \tag{6.4}$$

The F-measure determines the harmonic mean between *Precision* and *Recall* and is commonly used in place of accuracy as it provides a more detailed analysis of the data. Using this measure advances on just using accuracy for results in

TABLE 6.1: Results using *T1* and *T2* including Recall, Precision and F-measure on the SAMM dataset.

Method	Recall	Precision	F-Measure
LBP - [109]	0.5171	0.6084	0.5595
HOG - with *T1*	0.4657	0.7181	0.5650
LBP - with *T2*	0.7829	0.6508	0.7108
HOG - Proposed method	0.8429	0.7041	0.7672

Chapter 4. The equation can be defined as

$$F\text{-}Measure = \frac{2TP}{2TP + FP + FN}. \tag{6.5}$$

6.5.2 Peak Detection

As the peaks formed from the feature difference analysis are quite small, it was required to complete the peak detection manually. The process involved plotting each movement and thresholds and cross-checking with the ground truth FACS coding to determine a TP, FP or FN.

6.5.3 Method Comparisons

As the method described in Moilanen et al [109] is a similar feature difference method to the one proposed, it is used for comparing results. The original method is replicated on the SAMM dataset, and the threshold calculation (*T1*) is used for the first time with the spatial HOG feature. The baseline threshold (*T2*), is applied to both LBP and HOG.

The best result was Recall at 0.8428 using the proposed method and the *T2* threshold. This is much higher than the performance using LBP or HOG with the *T1* threshold with 0.5171 and 0.4657 respectively. Due to the dataset being high-speed videos and a higher resolution compared with previous experiments, it is likely that calculating distance will lead to large difference value peaks. *T1* was calculated using the mean and maximum values of the feature vector, so it will always detect a peak. This has the disadvantage in a real-world scenario if there are no movements (neutral face or no expression) they will always be misclassified as a movement.

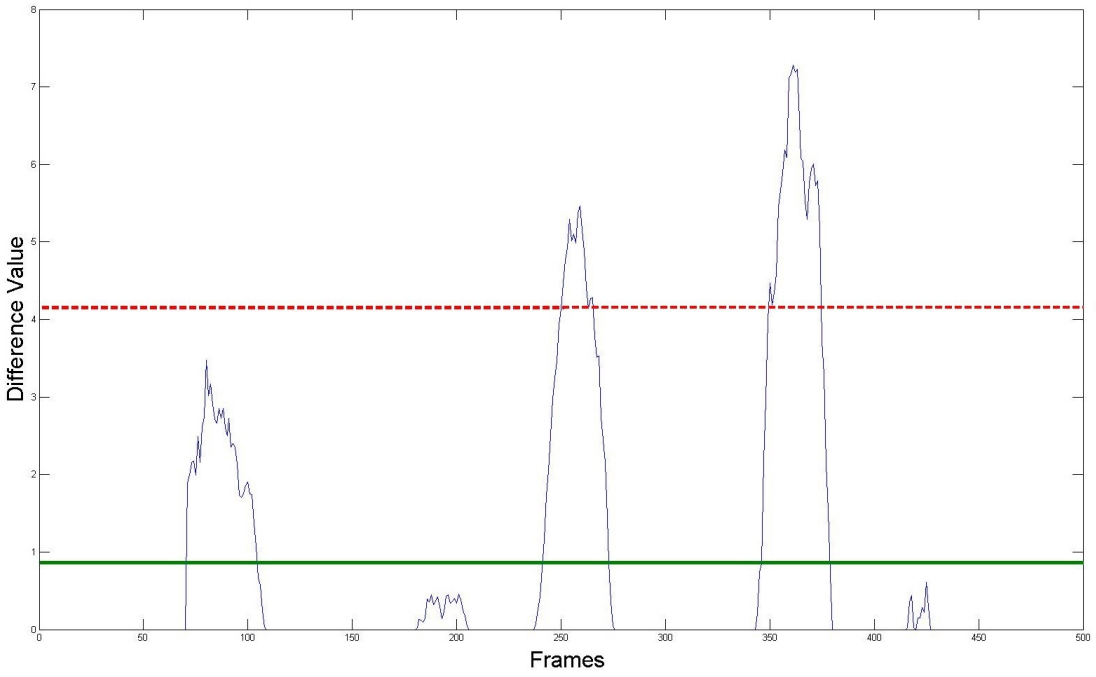

FIGURE 6.4: Illustration of a micro-movement sequence where a micro movement (peak 1) is missed by *T1*. The green solid line shows *T2*, which detects all micro-movements.

Using *T2*, the neutral expression for the participant is used as a threshold, therefore it is not affected by the values of the target sequence. Fig. 6.4 shows how a large peak increases the *T1*, whereas the proposed method detects the large peak but also peaks that have been missed due to using the sequence in the threshold value calculation.

In Fig. 6.5 *T1* uses the neutral sequence to calculate a threshold that appears to have detected movements, but in fact are the baseline neutral expression of that particular participant. Further, when a neutral sequence is input, the baseline stays above the peaks as they are below the participant's baseline.

Due to the proposed method's sensitivity, the Precision is affected and more FPs are observed. This can be seen in Table 6.1 as the Precision value between methods is relatively small, and using *T1* with HOG has a better Precision value of 0.7181 compared with using HOG and *T2*, which achieved 0.7041. Further tuning of the baseline selection would be required to reduce false positives but preserve the sensitivity of micro-movements detection. To provide a fair comparison with the proposed baseline method, only methods that have been used for micro-movement detection have been used, therefore LBP-TOP with GD proposed in Chapter 4 and was used for recognition, is not used for comparison in Table 6.1.

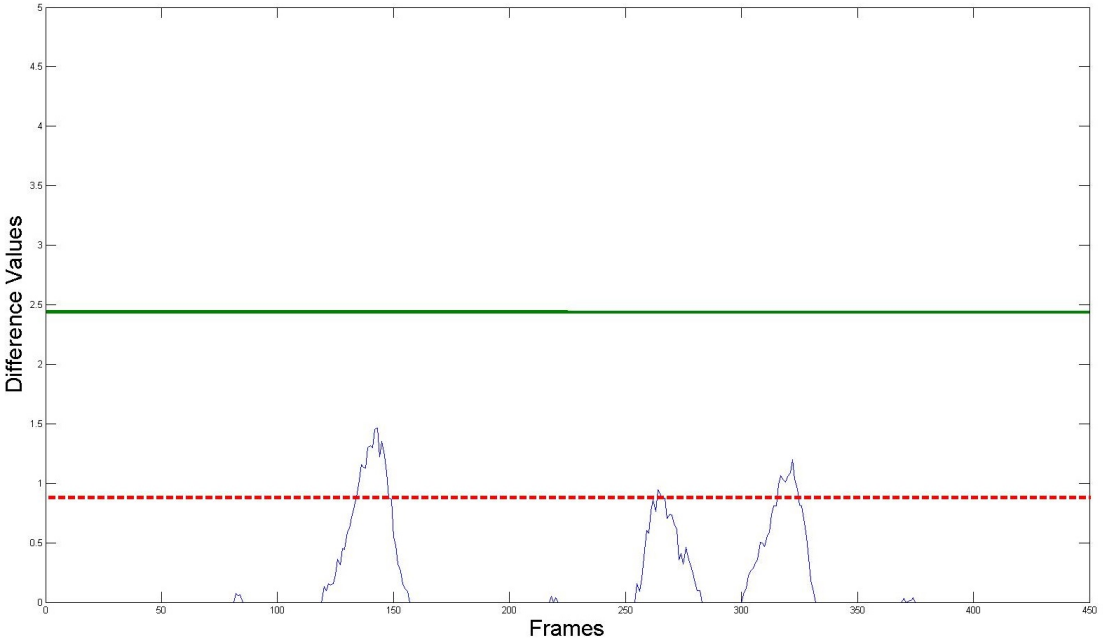

FIGURE 6.5: The green solid line shows that *T2* stays above values that are considered to be a person's baseline. The dashed red line shows peaks falsely detected by the previous detection method using *T1*.

TABLE 6.2: The best results of the previous feature difference method and the proposed ABT method, using the spatial HOG feature. The SAMM and CASME II datasets are compared.

Method	Dataset	Recall	Precision	F-Measure
Moilanen et al. [109]	SAMM	0.5171	0.6084	0.5595
Moilanen et al. [109]	CASME II	0.5219	0.2123	0.3031
Proposed Method - ABT	SAMM	0.9125	0.7304	0.8179
Proposed Method - ABT	CASME II	0.3951	0.2361	0.2961

6.5.4 ABT Results

The results presented in Table 6.2 show that the method in [109] does not perform as well on the SAMM dataset and mirrors a similar problem exhibited in their own results on the higher resolution CASME-A dataset, where the recall was 0.52. In the CASME II dataset, both methods performed poorly due to the same reasons, with the recall in [109] being similar at 0.5219. With the lower amount of frames available (i.e. fewer baseline frames) the N interval value had to be set to 13 for CASME II, and more noisy peaks were detected. As the SAMM dataset has the highest available resolution on micro-movements, it shows that the method in [109] struggles to process such data effectively.

By contrasting and comparing the baseline feature and movement feature using ABT, the proposed method substantially increases the detection rate and produced the best result of 0.9125 and 0.8179 for recall and F-measure, respectively. We observe that the precision has increased by 3% when compared to the results of the proposed method in Table 6.1. The proposed ABT method manages to reduce some of the FP, but overall still remains a challenge for micro-movement detection.

6.6 Discussion

Results from the original baseline outperformed the state of the art when comparing with Moilanen et al. [109]. However, further needs to be done to address limitations. Firstly, using only spatial (XY) features does not allow for temporal difference considerations, and usually leads to the lowest result overall [9].

Both proposed methods in this Chapter use a block-based approach, in other words, split the face into a set amount of blocks to extract features from. These blocks can also be describe as small video cubes, referencing the 3D video volume. One of the biggest limitation of this approach is that it can introduce unwanted descriptors from parts of the face that are not useful for micro-movement detection. Examples of irrelevant parts would be the neck and hair. A ways of localising the face parts to remove this information would be very beneficial for local feature analysis.

Results from the proposed ABT further outperformed the state of the art in both machine learning and difference analysis based approaches. Shreve et al. [6] obtained 74% TPs, 26% FNs and 44% FPs. Moilanen et al. [109] achieved the best result of 71% TPR. Finally, Li et al. [9] provided the most comparable results with a highest AUC of 92.98%. It should be noted that many results show different metrics for performance, indicating a need to standardise result outputs for a fairer comparison. Further experiments on other datasets than SAMM and CASME II would be advantageous to test the robustness of ABT, however the lack of baseline sequences within other datasets currently limit the experiments.

All results were calculated after manual peak detection completed by cross-checking when a movement peak crossed the threshold with the ground truth FACS coding. Obviously, this way of obtaining results would not be realistic for

a real-world scenario. One way around this problem would be to use automatics peak detection. As all the peaks are shaped similar to a Gaussian curve [150], if the zero-crossing is found and the threshold is crossed, then results could be generated much faster.

6.7 Summary

In this Chapter, a way of detecting micro-movements in the SAMM dataset was created by using participant neutral baselines as part of a threshold to determine when a micro-movement had occurred. An Adaptive Baseline Threshold (ABT) was introduced to balance the baseline feature values and adapt to the movement that was currently being processed.

To summarise, areas for improvement are detailed. Each dataset sequence used, whether this be a micro-movement or baseline sequence, was split into even blocks of 5×5 in dimension (25 blocks in total). Even though good results were obtained, splitting the whole face can include irrelevant information such as hair or have muscle regions split across multiple blocks.

It would be useful to know where on the face a person exhibits a micro-movement. The advantages of this include AU identification and to aid user understanding. By using a block-based approach and averaging the blocks with the greatest feature difference values, the local information about where on the face the movement occurs is lost.

Chapter 7

Local Feature Analysis with FACS-Based Regions

In this Chapter, 26 face regions are defined based on Facial Action Coding System. The proposed method becomes fully objective when a focus is placed on the muscle activation rather than emotional interpretation. Using these regions means that irrelevant information, such as hair, is removed during feature vector calculation, leading to better accuracy in determining a correct micro-movement. The method is validated on the two most recent micro-movement datasets: SAMM and CASME II.

7.1 Introduction

In this Chapter, 26 regions based on FACS are proposed to solve the problem of providing useful local information. Further, these regions are specifically created by FACS coders to align the regions to AUs, removing any irrelevant information that would be present in a block-based approach. The regions are fitted to the shape of each participant's face by using a PWA transform. The main method fits the mask to the face, however fitting the face to the shape of the mask is also discussed.

3D HOG [5, 98] is used as the feature that should best describe micro-movements, as the spatial only planes (XY) performed the best in the previous Chapter. Two other temporal features that have been used in micro-movement research,

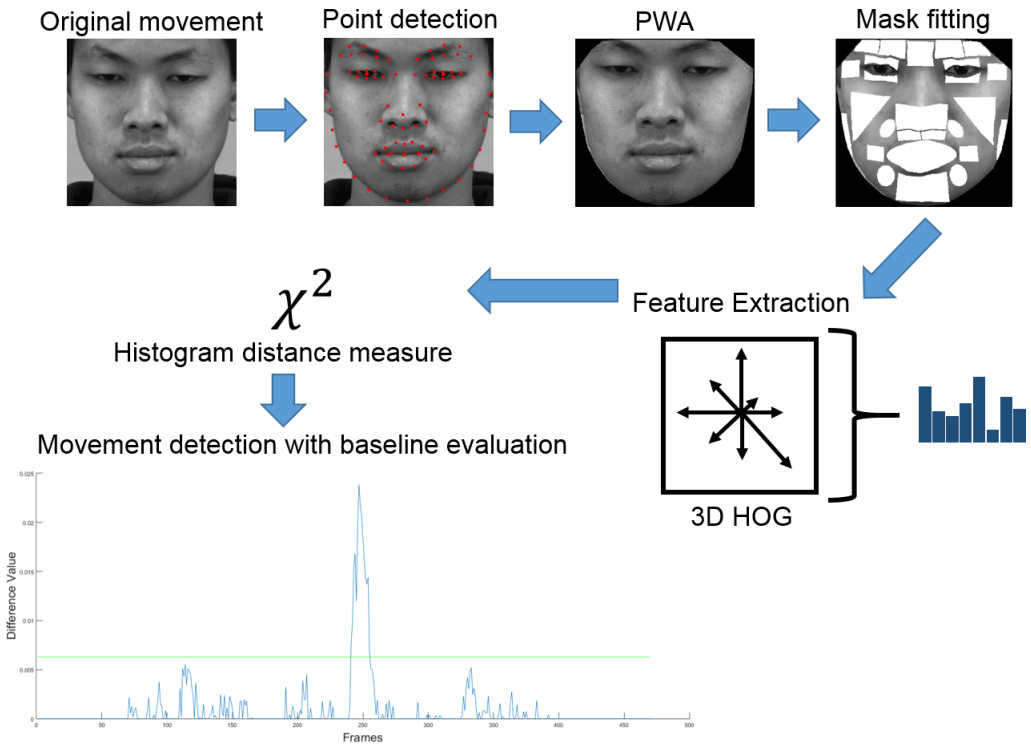

FIGURE 7.1: The processing pipeline of the proposed method using FACS-based regions for micro-movement detection.

LBP-TOP [17, 94] and HOOF [9, 145], will be used to test the robustness of this approach. The proposed SAMM dataset is used alongside CASME II [17] dataset to validate this approach.

This Chapter will conclude with an novel algorithm that combines the feature differences obtained from the local regions to output a video sequence with the local regions highlighted. Showing where a micro-movement occurred on the face can help the user understand micro-movements further, ready for interpretation, and allows for AUs to be predicted. The overall pipeline for this proposed method can be seen in Fig. 7.1.

7.2 FACS-Based Regions

To ensure only relevant movements are detected, 26 FACS-based [39] regions are proposed. Each region has been selected by three FACS-certified coders to focus only on areas of the face that contain movements from particular AUs. The advantage of this is that features can be locally analysed, the intensity of individual

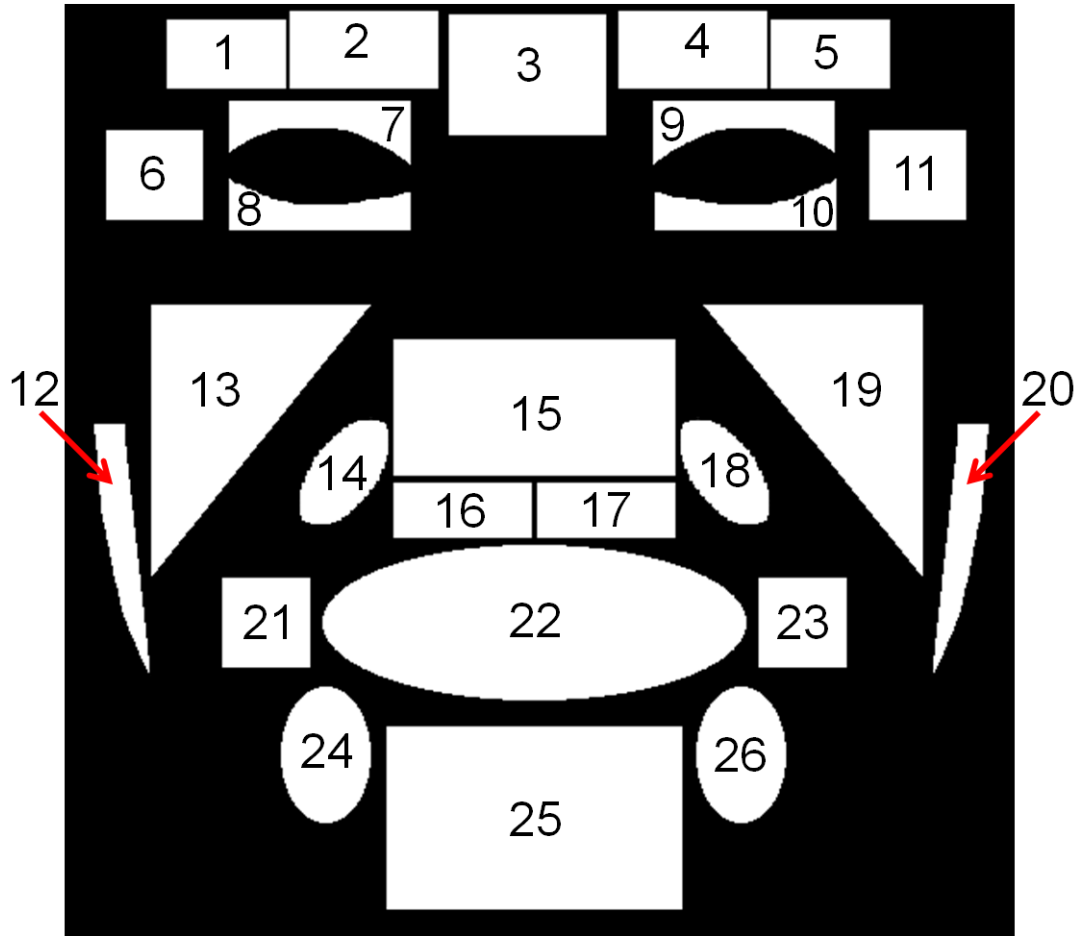

FIGURE 7.2: The original FACS-based region mask, with regions numbered from 1-26.

regions can be independently studied and not processing insignificant parts of the face (e.g. the hair).

It would be useful to know which regions correspond to which part of the face and which AU(s) are associated with that region. Fig. 7.2 shows the original face mask with each regions numbered from 1-26, and Table 7.1 summarises the name of each region and AUs that can occur in that area.

The process of using a PWA transform to fit the face regions as a binary mask to the face is summarised in Fig. 7.3. The method described warps the mask individually to each participant in a dataset based on the points obtained during the PWA transform (see Fig. 7.4). Even though the mask changes shape based on the person's facial features this does not affect movement classification as each participant in a dataset is handled individually rather than a set of abstract classes. When warping the mask to the individual face points, the different face

TABLE 7.1: A breakdown of the FACS-based region mask, showing the region numbers, names and associated AUs.

Region Number	Region Name	Associated AU(s)
1	Right Brow - Right	2, 4
2	Right Brow - Left	1, 4
3	Glabella	1, 4, 9
4	Left Brow - Right	1, 4
5	Left Brow - Left	2, 4
6	Crows Feet - Right	6
7	Upper Eye - Right	5
8	Lower Eye - Right	6, 7
9	Upper Eye - Left	5
10	Lower Eye - Left	6, 7
11	Crows Feet - Left	6
12	Jaw - Right	31
13	Cheek - Right	6, 12
14	Nasolabial - Right	11, 12, 13
15	Nose	9, 10, 38, 39
16	Upper Lip - Right	10
17	Upper Lip - Left	10
18	Nasolabial - Left	11, 12, 13
19	Cheek - Left	6, 12
20	Jaw - Left	31
21	Dimple - Right	12, 13, 14, 18, 20
22	Mouth	12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 28
23	Dimple - Left	12, 13, 14, 18, 20
24	Chin - Right	15, 17, 25, 26
25	Chin	15, 17, 25, 26
26	Chin - Left	15, 17, 25, 26

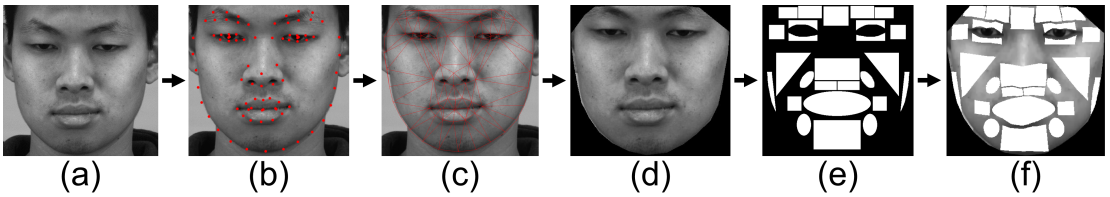

FIGURE 7.3: The process of piecewise affine warping for this method. (a) the original image (b) the original face has points automatically detected (c) Delaunay triangulation creates the convex hull for to allow the mask to be warped (d) the cropped face used for feature extraction and mask fitting (e) the original mask is warped to the shape of the cropped face (f) the final warped mask that has been applied to the cropped face.

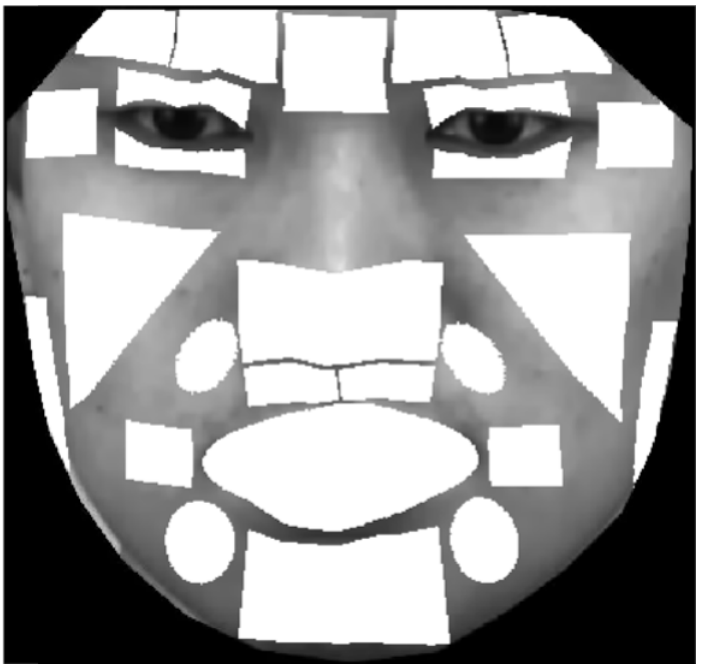

FIGURE 7.4: The face shape of each participant stays the same, with the FACS-based mask being warped using PWA transformation to each face. This creates a unique mask for each person.

proportions are taken into account to avoid personal face differences. PWA was outlined by Cootes and Taylor [136] and is described in Section 3.1.4.

The other option is to warp the face to a static mask shape, however this leads to faces becoming distorted and potentially removing any micro-movements. One example of a problem arising by warping the face to the mask is that regions may not fit to the best position. This includes when a person is wearing glasses or have facial features that are much different to usual, such as a very large nose. An example of warping the face to fit the mask can be seen in Fig. 7.5.

7.3 Micro-Movement Region Localisation

To represent the movements for analysis, 3D HOG is used for the first time in micro-movement detection. Initial processing includes de-noising to reduce high-speed video noise. Three planes (XY, XT and YT) are extracted using 3D HOG to describe different directions of motion. Two additional plane representations are created by concatenating the XY, XT and YT planes and the temporal XT and YT planes.

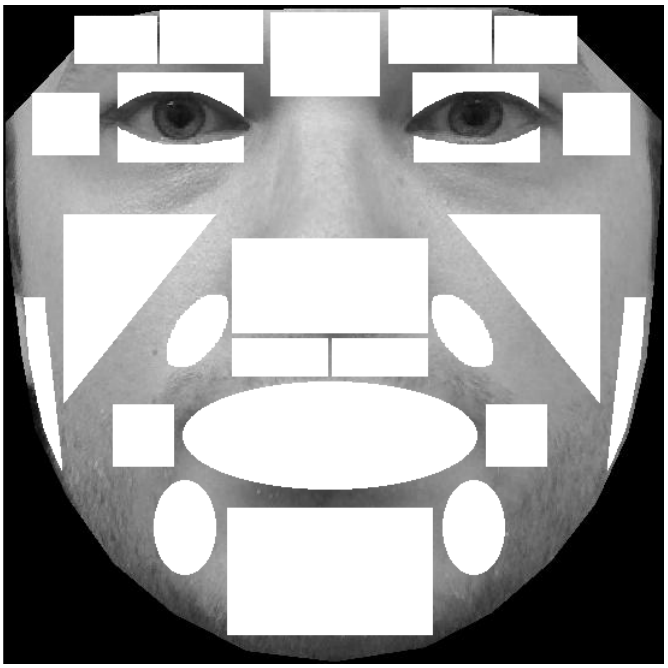

FIGURE 7.5: The FACS-based regions stay the same shape as the original mask, and the face is warped using a PWA transform to fit this shape.

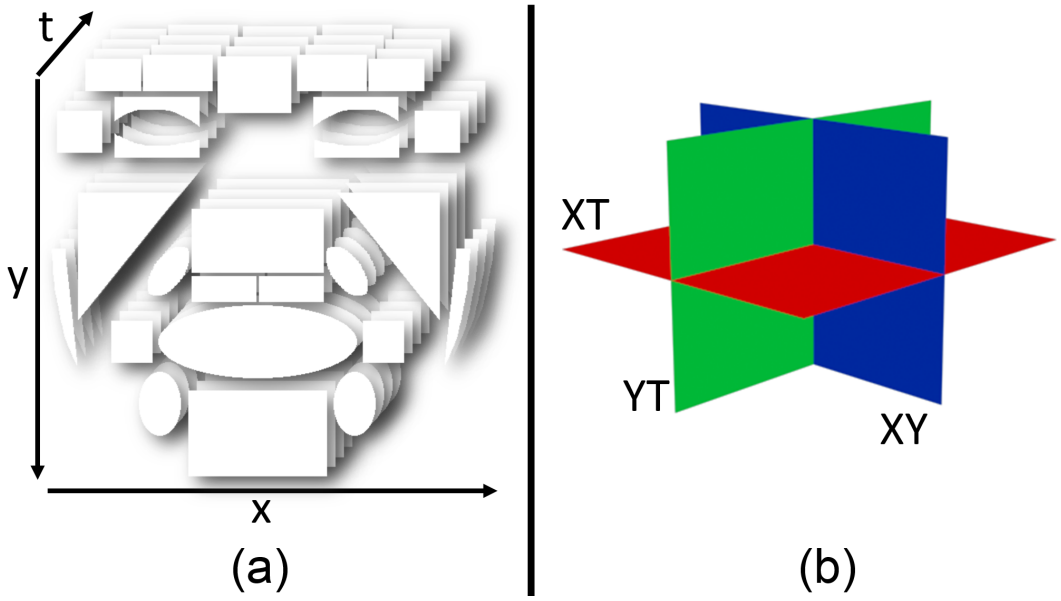

FIGURE 7.6: (a) Visual representation of the spatio-temporal configuration of video frames split into blocks. (b) The XY, XT, YT planes used for feature analysis in LBP-TOP and 3D HOG.

It is only natural to extend HOG into temporal when analysing video data. Fig. 7.6(a) shows regions as a 'video cube' that contain each frame of the video cropped in that particular region. Fig. 7.6(b) shows a representation of three orthogonal planes in 3D space. Each plane outlines the axis which is being analysed in the video - XY, XT and YT.

For both SAMM and CASME II, feature extraction is used to represent the movements and baselines defined by the FACS coding ground truth. Each movement start and end point has the 3D HOG feature applied to create temporal representation in each of the defined regions. The extracted movement also includes neutral frames at the start and end of the sequence so the movement has an onset and offset point.

The baseline feature is extracted similarly to the movements, but uses a set amount of frames to define a subject's baseline. The features are once again split into regions to represent the baseline of that particular local area for that particular person. The amount of frames used for SAMM was 200, whereas 50 was used for CASME II due to the lack of available neutral frames.

7.4 Micro-Facial Movement Detection

In the proposed method, micro-movements are treated objectively as facial muscle activations, like in FACS, and the temporal differences are calculated using the χ^2 distance. Previous difference methods [6, 93, 109] all assume that the detected movements in the videos are actual micro-expressions. The datasets they use can sometimes confirm this, but it creates a very constrained analysis. The proposed method uses a baseline expression vector calculated from the neutral face sequence of participants to create individual baseline thresholds to determine what is a micro-movement for that person.

The distance function described derives from Moilanen et al. [109] but extends to the temporal domain, uses FACS-based regions with 3D HOG, includes region normalisation and baseline thresholding using ABT. Finally, peak detection is used to automatically find movement peaks and output the onset, apex and offset frames of each movement rather than relying on arduous manual annotation.

As in [109], the k-th frame is described as

$$k = \frac{1}{2}(N - 1) \tag{7.1}$$

where N is odd numbered micro-interval value. As the SAMM dataset was recorded at 200 fps, it is calculated that the value of N should be around 71. The value of k in this instance would be 35.

The difference between the current frame and average feature frame shows facial changes in a particular region. The possible change in the features is rapid since it occurs between start frame and end frame, to distinguish the quick changes from temporally longer events. The difference analysis continues for each frame except the first and last k frames that would exceed the boundaries of the video. This also means rapid facial movements such as blinks would also be classified as a movement, and so the proposed region mask removes as much of the eye as possible without removing the important muscle areas around the eyes. The χ^2 distance is used for histogram difference analysis and is defined as

$$\chi^2(P,Q) = \sum_{b}^{B} \frac{(P_b - Q_b)^2}{(P_b + Q_b)} \tag{7.2}$$

where b is the b-th bin in the P and Q histograms that have an equal number of bins for a total amount of bins B. All temporal planes (XY, XT, YT) are used in varying combinations, for example, concatenating the XT and YT planes to form the XTYT feature. Finding the dissimilarity between these features investigates what is more suitable to represent micro-movements.

Unlike previous methods that split the images into even-sized blocks, the regions used in this method are all different sizes. This leads to the issue of the larger regions having more pixels in the area of the mask and therefore differences that are not reflective of which regions is more meaningful. The normalisation is defined as

$$\mathbf{F}_i = \frac{\mathbf{F}_i}{A_r} \tag{7.3}$$

where A_r is the number of pixels in each individual binary bask region and i is the index of the value within the feature $\mathbf{F}$. This step is completed for each region and was done to make sure that the area of regions that were not rectangular in a matrix were still accurately calculated to normalise the regions correctly.

There is an option to select the top regions, defined as R, with the greatest difference values to form the first feature vector, $\mathbf{F}_{R,i}$, that represents the overall movement of the face and is defined as

$$\mathbf{F}_{R,i} = \sum_{r=1}^{R} (D_{r,1}, D_{r,2}, \ldots, D_{r,i}) \tag{7.4}$$

where D is the difference values of each individual region, j, sorted in descending

order up to R for each frame, with i being the total number of frames or index value of that feature.

For the proposed method, each region feature is calculated initially and ranked in descending order from highest difference value to lowest. R regions are then sampled from $(2, 4, 6, \ldots, 26)$. In other words, with the highest difference, we iterate by adding two regions until total amount of the proposed regions (26) are used. This then forms the Receiver Operating Characteristic (ROC) curves for the results from each feature. The best performing amount of regions was determined to be 12 using these tests, and the results in the next Section show this.

The average of the surrounding start frame and end frame values are subtracted from each value in the current frame for each region, R. Each new value of the i-th feature of region R is therefore calculated as

$$\mathbf{F}'_{R,i} = \mathbf{F}_{R,i} - \frac{1}{2}(\mathbf{F}_{R,i+k} + \mathbf{F}_{R,i-k}) \tag{7.5}$$

with each frame having Eq. 7.5 applied, apart from the first and last k frames of the each sequence due to the temporal boundaries. Any negative values in $\mathbf{F}'_{R,i}$ are set to zero, as any of these values indicated the value at the current frame was below the average difference values of start frame and end frame. These values represent no fast changes, and so are removed to avoid irrelevant data.

7.5 Experimental Results

The proposed method of using FACS-based regions with individualised baselines performs well on both SAMM and CASME II compared with the previous state of the art. Three other feature descriptors that have been used in micro-movement detection are implemented to compare with 3D HOG: LBP-TOP, HOOF and the Main Directional Mean Optical-flow (MDMO).

Automatic peak detection [150] was implemented to check that the apex of a detected peak was above the ABT and within the duration of the micro-movement that had been labelled during FACS coding. The ground truth was set for each region, where the ground truth AUs corresponded to regions of movement. By

TABLE 7.2: Performance analysis metrics for each feature descriptor on both datasets.

	SAMM			CASME II		
Feature	Recall	Precision	F-Measure	Recall	Precision	F-Measure
3D HOG - XY	0.5607	0.2831	0.3763	0.5134	0.4216	0.4631
3D HOG - XT	**0.6804**	0.3198	0.4352	0.6235	**0.5341**	**0.5754**
3D HOG - YT	0.5821	0.3043	0.3998	0.4857	0.4561	0.4705
3D HOG - XTYT	0.6661	0.3179	0.4305	0.6202	0.5157	0.5631
3D HOG - All Planes	0.6607	0.3181	0.4295	0.6202	0.5164	0.5636
LBP-TOP - XY	0.3732	0.2735	0.3157	0.2807	0.2427	0.2603
LBP-TOP - XT	0.35	0.2688	0.3041	0.3513	0.2876	0.3163
LBP-TOP - YT	0.3196	0.2716	0.2937	0.3101	0.2622	0.2842
LBP-TOP - XTYT	0.3464	0.2755	0.3070	0.3319	0.2714	0.2987
LBP-TOP - All Planes	0.3482	0.2589	0.2970	0.3134	0.2560	0.2818
HOOF	0.4214	0.2486	0.3128	0.4723	0.3351	0.3920

TABLE 7.3: The AUC values for each feature descriptor on both datasets.

Feature	SAMM	CASME II
3D HOG - XY	0.6910	0.6454
3D HOG - XT	**0.7513**	**0.7261**
3D HOG - YT	0.7278	0.6486
3D HOG - XTYT	0.7513	0.7157
3D HOG - All Planes	0.7481	0.7183
LBP-TOP - XY	**0.6401**	0.4631
LBP-TOP - XT	0.6197	0.4853
LBP-TOP - YT	0.6173	**0.4866**
LBP-TOP - XTYT	0.6144	0.4747
LBP-TOP - All Planes	0.6154	0.4623
HOOF	**0.6312**	**0.5203**

having the ground truth labelling for every region in each movement, it was possible to extract the common True Positives (TPs), False Positives (FPs), False Negatives (FNs) and the less common True Negatives (TNs).

Firstly, the ROC curves are presented for three types of feature descriptors on the SAMM (Fig. 7.7) and CASME II (Fig. 7.8) that plots the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis.

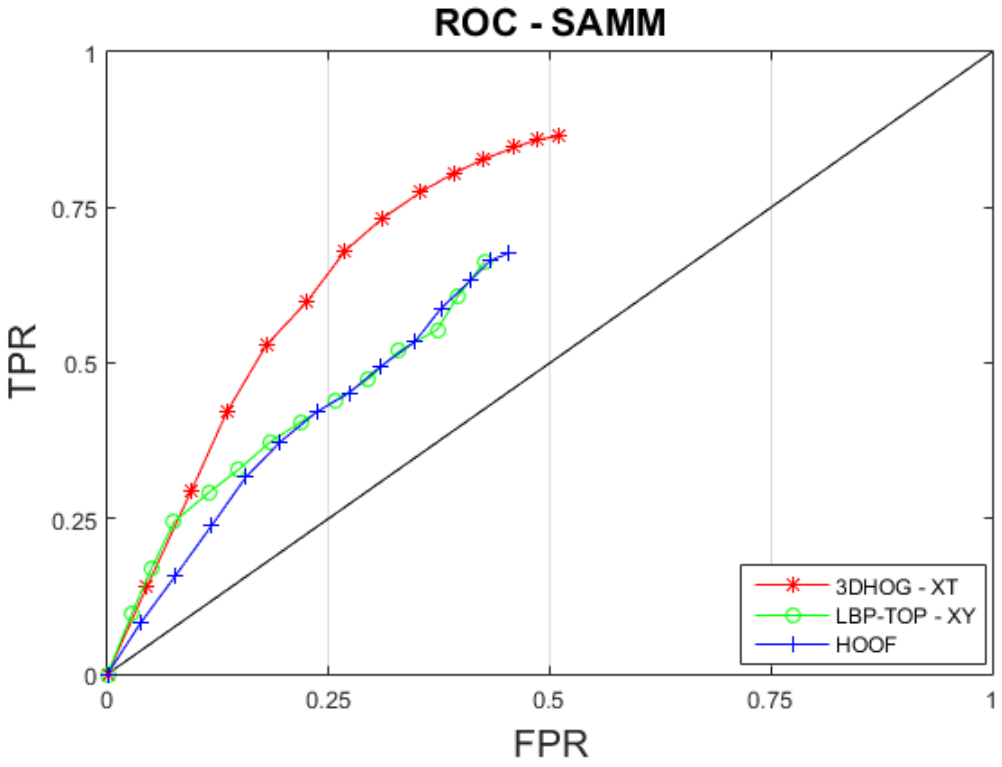

FIGURE 7.7: Three ROC curves shown for the SAMM dataset. Each descriptor corresponds to a different curve with 3D HOG performing best.

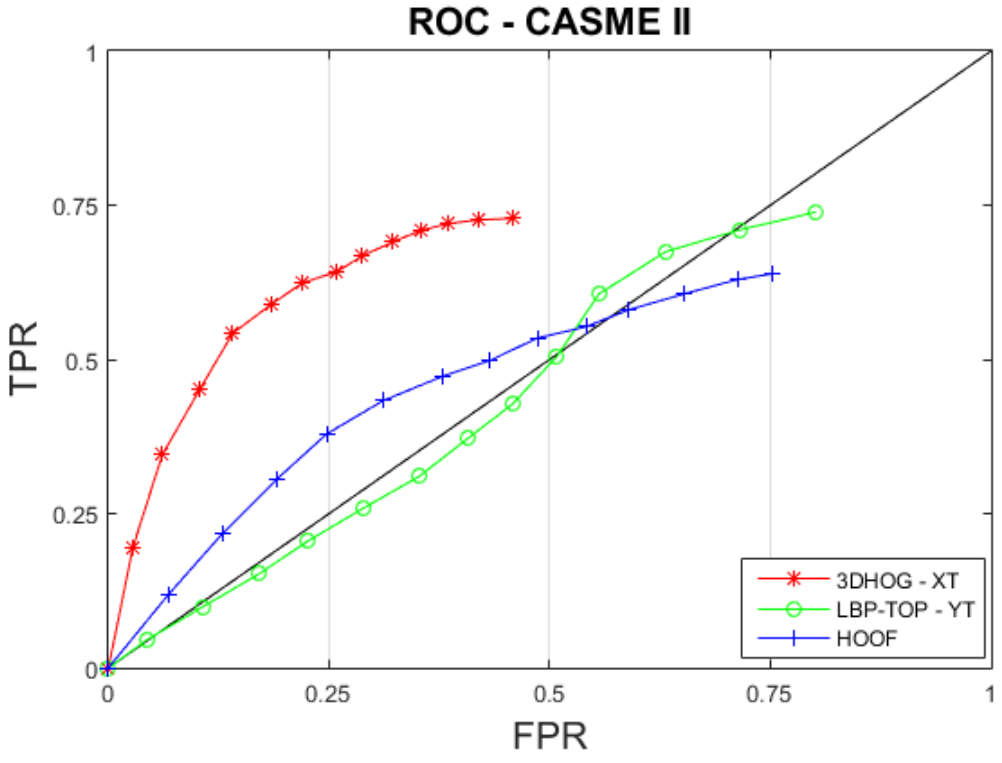

FIGURE 7.8: Three ROC curves shown for the CASME II dataset. Each descriptor corresponds to a different curve with 3D HOG performing best.

TABLE 7.4: A comparison of the current micro-movement methods. Each result metric changes depending on the method. * = true positives/recall, ** = area under curve.

Method	Feature	Dataset	Result
Moilanen et al. [109]	LBP	CASME II/SMIC	71%*
Shreve et al. [6]	Optical Strain	USF-HD	74%*
Li et al. [9]	LBP	CASME II	92.98%**
Xia et al. [112]	ASM Geometric Deformation	CASME/CASME	92.08%*
Patel et al. [113]	Optical Flow	SMIC	95%**
Proposed - ABT	LBP/HOG	SAMM	91.25%*
Proposed - FACS-Regions	3D HOG	CASME II/SAMM	68.04%*

The points of the ROC curves are calculated by sampling R regions from 2 until the total of R,$(2, 4, 6, \ldots, 26)$. The highest difference values are used for the feature descriptor up until all R regions are used. As the curves show, as the more regions are introduced, both TPR and FPR increases as predicted when more regions become available for detection. Further, the plots show that 3D HOG outperforms LBP-TOP and HOOF on both datasets.

In addition to the ROC curves, the Area Under Curve (AUC) is calculated for each feature in both datasets, shown in Table 7.3. 3D HOG performs best on the SAMM and CASME II dataset in the XT plane with 0.7513 and 0.7261 respectively. LBP-TOP achieves 0.6401 in the XY planes for the SAMM dataset, however it does not perform as well in the CASME II dataset with the highest result in the YT plane with 0.4866.

The performance analysis scores of recall (TPR), precision and F-measure are presented in Table 7.2. Each score was selected from the number of top regions that performed best in describing the micro-movements, which in these experiments was 12 regions. Overall the best feature descriptor for SAMM was 3D HOG in the XT plane achieving 0.6804, 0.5341 and 0.5754 for recall, precision and F-measure respectively. Table 7.4 shows a summary of micro-movement detection methods with their results. This method performs well and is comparable with the micro-movement detection results in [9, 109]. However, the precision is low on the majority of cases due to the larger amount of false positives detected. Unfortunately, as with any recordings of humans in video sequences, it is difficult to

eliminate all head movements, even with face alignment. As micro-movements are subtle motions, any head movements are likely to contribute to the final difference values, therefore skewing the accuracy of spotting real micro-movements.

The higher precision seen in CASME II compared with SAMM can be explained by the lower number of frames that were available to analyse in the CASME II dataset, meaning FPs are less likely and spotting is easier. The original recordings of the CASME II participants are no longer available, and so creating a longer sequence was not possible at this time.

Higher spotting accuracy can be achieved through the increase of regions, however this tips the balance of TPs and FPs, lowering the overall F-measure. For example, using the 12 regions described in Table 7.2 the 3D HOG - XT feature on SAMM the recall is 0.6804, however if this is increased to 26 regions the recall score jumps to 0.8643 but is accompanied by a lower F-measure score of 0.3752. The use of ranking regions to balance the trade off between detecting more movements and to exhibit less sensitivity, was proposed in Moilanen et al. [109] and advanced upon in Chapter 6 to remove features that may introduce noise and redundant information.

This method takes away any assumption based on emotions and reduces the movements to their basic definition of muscle movements. Doing this allows for further interpretation at a later stage rather than attempting to use techniques, such as machine learning, to define micro-movements into discrete categories.

7.6 Applications

During development of novel micro-movement methods, the discussion of real-world applications was ongoing. From the research findings in this work, a few points are highlighted to discuss the potential of applying what has been found to real-world problems.

7.6.1 Frame Rate Sub-Sampling

In most of the experiments, the frame rate has set to be 200 fps. However, processing this many frames in a second interval is very expensive in terms of computation

time. As real-world applications would require a system to run in real-time, a lower frame rate would be required. An experiment was run to test the best performing feature, 3D HOG in the XT plane, on the SAMM dataset (see Table 7.2) with the frame rate sub-sampled to 100 fps.

The number of regions chosen for the final descriptor results was 12, as with previous results. It was found that the recall with sub-sampling was 0.75, the precision was 0.42 and the F-measure score was 0.37. These results outperform the recall and precision of the 200 fps experiment, however the F-measure score was lower by around 6%. The higher results on the first two metrics is likely due to the half the amount of frames to be checked, which can also be reflected on the lower overall F-measure.

Although the increased results may be due to the lower frame rate, it could be possibly to have a trade-off between accuracy and computational complexity. For example, the detection accuracy increases, but it may detect more false positives. However, the time to detect micro-movement has decreased as half the amount of frames need to be processed.

7.6.2 Movement Localisation

An improvement from using a block-based approach in Chapter 6 is to create novel FACS-based regions that represent local areas of the face that correspond to specific facial muscle actions. Each of the regions have an individual feature difference value, allowing for each region to have different magnitude changes. Fig. 7.9 shows a micro-movement of a participant from SAMM, with regions identified using highlights when movement occurs.

When the highlight of the identified region becomes more opaque, the movement is larger and vice versa, therefore allowing for the representation of movement intensity. This feature is useful for pinpointing where on the face the largest feature difference has occurred. However, on its own, this feature is limited in providing feedback to the user, but is the first step towards a real-world application.

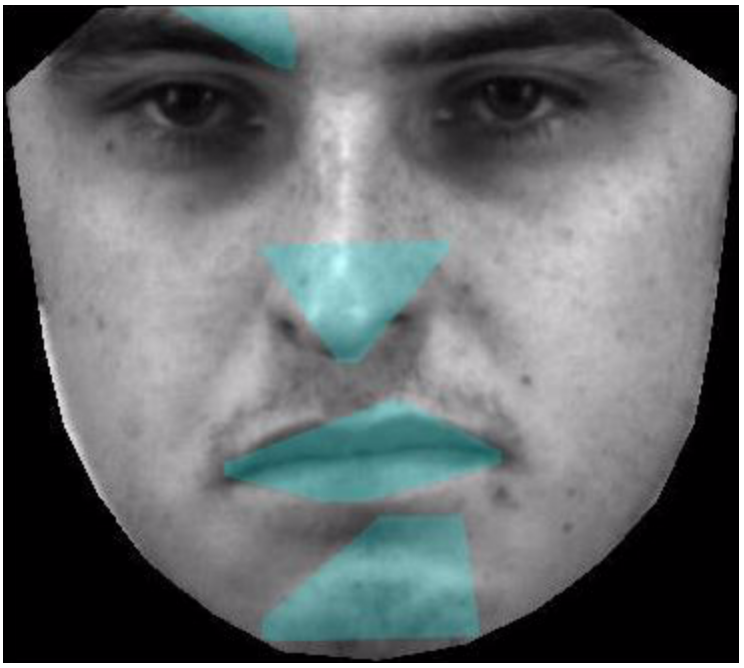

FIGURE 7.9: A micro-movement from SAMM showing regions highlighted when movement occurs in that area.

7.7 Summary

This Chapter has outlined new FACS-based regions created to localise feature descriptors to parts of the face that contain muscle movements, removing the irrelevant parts of the face from the final feature vector. Further, by using 3D HOG the local regions are tested against two other temporal features: LBP-TOP and HOOF. These features are then tested on two datasets, SAMM and CASME II, as the most recent datasets available and are comparable in terms of data content.

Currently, the SAMM dataset is the only known dataset to provide enough data to produce baseline sequences as other micro-expression datasets, such as SMIC [15], CASME [16] and CASME II [17], do not have the raw captured sequences available for similar use. The CASME II baseline sequences in this method had to be obtained through the excess frames provided after a movement had occurred. To thoroughly compare baseline methods with SAMM, further data collection of FACS-coded spontaneous micro-movements is required.

As a proof of concept, a micro-movement detection prototype was created to test it's potential for real-world use. It was found to be too computationally intensive to be classed as real-time, and further work would be required to optimise

this system. In addition, other applications that this method could apply to are discussed as future research and development projects.

Chapter 8

Conclusion

In this final Chapter, a summary of the contributions of this thesis on micro-movement detection are discussed. A critical analysis of the work completed is done with a focus on the strengths and limitations found during the research. It also highlights potential future improvements to this field and the direction in which it is heading for researchers in this continuously growing area.

8.1 Introduction

Micro-facial movements are a constant subject of debate in psychology, especially in the area of deception cues. Using computer vision to detect these subtle changes brings together historical knowledge of micro-movements and technology to create a way of automatically detecting micro-facial movements. However, there are still many challenges to face, including how best to represent micro-movements as feature descriptors and how to differentiate between a real micro-movement and other non-relevant motion. A novel feature and dataset was created to form foundations for further study. Research then moved into more recent developments, inspired by [109], on feature difference analysis of micro-movement. This objective method of plotting micro-movements as magnitude peaks leads to the study on human baselines as features, and local FACS-based regions.

This thesis proposed novel methods to advance the field of micro-movement detection: the first being an investigation into the recognition of micro-movements

using a new descriptor combining LBP-TOP and GD. Second, a new micro-movement dataset, Spontaneous Activity of Micro-Movements (SAMM), is created to add to the limited datasets available. Next, micro-movement detection moved towards a feature difference approach, taking a more objective route to solve the problem. In addition, by using the new SAMM dataset a novel individualised baseline feature was created to find a threshold that defined a micro-movement. Finally, new FACS-based regions were proposed as localised feature descriptors that processed only the areas of the face that was relevant to FACS.

8.2 Research Findings

A summary of the research objectives is shown in Table 8.1 along with the corresponding outcomes. These findings will detail the reason for each objective and how the outcome was achieved.

The first objective is to investigate the potential of being able to recognise micro-movements from non-movements using temporal features and machine learning algorithms. This was achieved by proposing LBP-TOP with GD (see 4.3), and testing how micro-movements are recognised using the machine learning algorithms SVM and RF. LBP-TOP has already been used in the classification of micro-expressions [17], however by combining it with GD some feature such as lines and blobs of the face were accentuated, giving a more pronounced feature for describing subtle movements.

The CASME II dataset [17] was used as a benchmark dataset due to it being the most recent micro-movement dataset available and contained the largest amount of FACS coded micro-expressions. The primary goal in Section 4.4 was to investigate how micro-movements and non-movements would be classified using two popular machine learning methods: SVM and RF. The dataset was able to be split into testing and training data based on the ground truth provided. The results were promising, achieving the highest accuracy score of 92.6% using RF and a 50:50 ratio of testing and training data. However, the results only reported accuracy, rather than any other form of performance measure (i.e. F-measure), meaning false positives and false negatives were not used effectively. Further, the results on SVM performed poorly even as training data increased. It was concluded that using machine learning is highly unpredictable, and being able to find

TABLE 8.1: The research objectives (defined in Section 1.4) against the actual outcomes.

No.	Objective	Outcome
1	To investigate the potential of being able to recognise micro-movements from non-movements using temporal features and machine learning algorithms.	An extended LBP-TOP feature combined with GD is created and tested using SVM and RF.
2	To create a new spontaneous micro-movement dataset by conducting an emotional inducement experiment.	A new spontaneous micro-facial movement dataset, SAMM, is created to address the low number of currently available datasets and their limitations.
3	To explore and compare different feature descriptors that best represent micro-movements for accurate detection.	The most recent features used for micro-movement detection were explored and an individualised baseline method using HOG features and a temporal difference method is proposed for micro-movement detection.
4	To propose an objective method of detecting micro-facial movements using localised features.	A novel micro-movement detection method based on FACS-based regions and individualised baselines with 3D HOG features was introduced.
5	To evaluate the performance of the proposed methods against benchmark algorithms and datasets.	The proposed methods and benchmark algorithms are evaluated on the SAMM and CASME II datasets.

a split between what is a micro-movement and a non-movement effectively, is unrealistic. It was also realised that a lack of datasets contributed to not being able to test the validity of methods.

The second objective is to explore and compare different features that best represent micro-movements for accurate detection. This was achieved by extending LBP-TOP with GD to form a new feature, and seven other features that have been used recently in the field of micro-movement analysis: LBP, HOG, LBP-TOP, 3D HOG, optical flow, HOOF and MDMO. LBP and HOG were first

used in Section 6.2.2 as spatial only features used in feature difference analysis. Using the spatial (XY plane) only version of these features allowed for replicated of the Moilanen et al. [109] results, and made way for expanding into the temporal domain. This was done using the remaining features, which were used in Section 7.3 to describe 3D HOG, the main feature used in the proposed FACS-based region method. The rest were summarised in the results in Section 7.5, where they were used to compare against 3D HOG to find out which feature outperformed the rest in micro-movement detection.

The third objective required an emotional inducement experiment to be completed with volunteers to create a new spontaneous micro-movement dataset. This was achieved by proposing a new spontaneous micro-facial movement datasets called SAMM. The dataset was create to address many of the limitations of current publicly available datasets, some of which included using participants from limited ethnicities, the mean age of participants was low due to students being used in many datasets, the relatively low resolutions used compared with the available technology and few are FACS coded. As part of the experimental protocols, participants were asked to fill in a questionnaire that detailed their personal emotional triggers. From these answers it was possible to tailor all stimuli to each participant and increase the changes of a response, especially when the instructions stated to try and hide their true emotion by keep a neutral face. Some limitations did arise, including the huge amount of data produced from recording at 200 fps for minutes at a time. This led to an external storage requirement to store over 5TB of participant recordings. Further, it would have been beneficial to gain information from participants on how they felt after every stimulus, so that more in depth statistical tests could be performed.

The fourth objective was to explore the ways in which the psychological research and computer vision can integrate and form novel methods of micro-movement detection. This was achieved by creating a novel method described in Section 6.3, an individualised baseline, resulting in an Adaptive Baseline Threshold (ABT). Knowing the facial muscle baseline of the participant is vital to be able to know from where a movement should be calculated from. By using an individualised baseline feature, the ABT can be calculated and adapt based on each participant's movements. The threshold can then be used to determine the point at which the feature difference magnitude values should cross to be counted as a micro-facial movement. The results were calculated by splitting the face in 5×5

blocks and by using HOG and LBP features on the SAMM dataset for the first time. Results for the proposed individualised baseline was a recall score of 0.8429 using HOG. With the introduction of the novel ABT, results were increased to a recall score of 0.9125.

Finally, the fifth objective was to evaluate proposed methods against benchmark algorithms and datasets to increase knowledge in this emerging field and to aid human interpretation. This was partly achieved in Section 7.4, with the creation of FACS-based regions that focus on local areas relevant to muscle movements. Created by certified FACS coders, the regions pinpoint facial areas that are relevant to muscle movement, and so in using these areas the method becomes completely objective. Further, by only using defined regions the irrelevant information on the face caused by using a block-based method, such as hair or the neck, are removed. The eyes are also not included in the calculations, however as the eyelid is so close to the upper and lower eye regions, some eye movement and blinking movements added to the feature vector as noise. The rest of the objective was achieved by validating the new method on two state of the art datasets: SAMM and CASME II. The highest result for SAMM was a recall score of 0.6804 and for the CASME II dataset it was a recall score of 0.6235. Both of these results used the 3D HOG feature on the XT plane. Overall it can be concluded that from these experiments, 3D HOG outperforms LBP-TOP and HOOF in micro-movement detection.

Although all objectives were met, the limitations of the research in this thesis should be discussed. First, the computational cost should be more clearly defined. To allow for real-time processing, computationally intensive techniques like HOG need to be optimised or replaced with more efficient techniques. Second, all experiments and data currently only deal with frontal face image, which are common in an experimental setting but not in a real-world setting. Finding micro-movements that occur during faces that are rotated or in more natural poses would increase content quality of current datasets. Third, the need for neutral sequences to find baselines is similar to requiring training data in traditional machine learning methods. Ways around this would be to find a baseline during a person's live recording and using this to define a threshold. Another method would record the baseline first before proceeding with micro-movement detection algorithms. However, in real-world environments it would not be possible to ask a person to keep a neutral face, even for a few seconds.

8.3 Future Work

The future of this field has a lot of potential to continue developing new solutions and further the knowledge and understanding. The following points are a few of the possible directions to take to advance this field.

8.3.1 Cross-Cultural Analysis

As first discussed in Section 2.5, emotional suppression across cultures has been comprehensively analysed [23, 47–52]. Micro-movements occur when people attempt to hide their true emotion, and so the possibility of how well some cultures manage this suppression would be interesting to learn. By using software to detect micro-movements across cultures, the results of different suppression of emotion can be studied. Therefore people in East Asian cultures, where emotional suppression is encouraged, and Western cultures, who do not encourage suppression [24], can be analysed to find any correlation between the psychological studies and micro-movements. Something to note in this type of investigation would be to ensure the different participants originate and live in their respective countries, as people living with different cultures for a long time may not exhibit the same behaviour.

8.3.2 Dataset Improvements

Even with the addition of the SAMM dataset, further work can be done to improve micro-movement datasets. Firstly, more datasets or expanding previous sets would be a simple improvement that can help move the research forward faster. Secondly, a standard procedure on how to maximise the amount of micro-movements induced spontaneously in laboratory controlled experiments would be beneficial. If collaboration between established datasets and researchers from psychology occurred, dataset creation would be more consistent.

As using human participants is required, and emotions are induced, ethical concerns are always going to play a part in future studies of this kind. Any work moving forward must take into account these concerns and draw from previous

experiments to ensure no harm will come to the psychological welfare of participants. The SAMM dataset (discussed in Chapter 5) attempts to set a precedent for following correct ethical guidelines.

8.3.3 Real-Time Micro-Movement Detection

To be able to implement any form of micro-movement detection system into a real-world scenario, it must perform the processes required in real-time (or near to real-time). As the accuracy of facial expression analysis is already quite high, transitioning to real-time has already produced decent results and methods [2, 167–169], however there is currently no known systems that is able to detect micro-movements. Further, the accuracy of many state-of-the-art methods is still too low to be deployed effectively in a real-world environment.

With increasingly better technologies and processing power, it may not be long before a real-time system is possible. The challenges would be how to track the face continuously, take in features, process using the detection algorithm, and output to the user all in less than 5 seconds (if not less). Doing all this would require a focus on parallel computing, where processes are split between processing cores to perform multiple tasks at once. The software would also rely on coding in a language such as C++, as the requirement is on the speed of completing the computing tasks.

8.4 Concluding Remarks

The work presented in this thesis forms contributions in micro-movement detection, where a focus is placed on ensuring the movement are treated objectively, rather than jumping to conclusions. By outperforming many of the state-of-the-art methods, the contributions show promising results for the future of this field, where it can only get better as knowledge and understanding increases. This field is a challenging topic, and in some ways can feel almost impossible to detect changes even a human struggles to see. However, as seen in the literature and work presented, micro-movement detection may soon be in the mainstream like facial expressions.

Appendix A

Haar Feature Calculation

Firstly, the Haar features are determined to outline what is required in an image for such a feature to exist. For a Haar feature to be present, the average pixel values of the dark region of pixels is subtracted from the opposite average of light region of pixels (see Fig. A.1). If the threshold that was set during learning is exceeded, then the feature exists in the image. A 24×24 pixel window is used to scan across the image for feature detection when comparing to the trained model. As an image may potentially have hundreds of features present, Viola and Jones [58] use an 'Integral Image', or Summed Area Table (SAT) to sum the dark and light region of pixels of features (Fig. A.1). In [58] the integral image at (x, y)

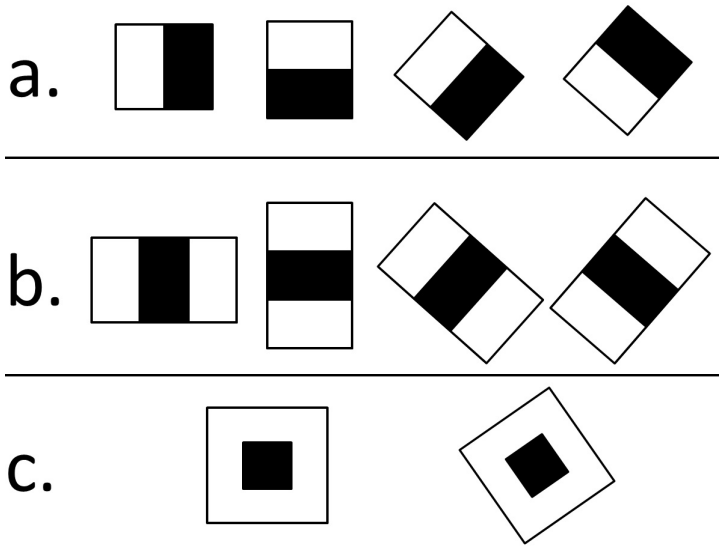

FIGURE A.1: The Haar features defined by [59] as an extension of the original features defined by Viola-Jones. The first features *a.* are edge features, *b.* are line features and *c.* are centre-surround features.

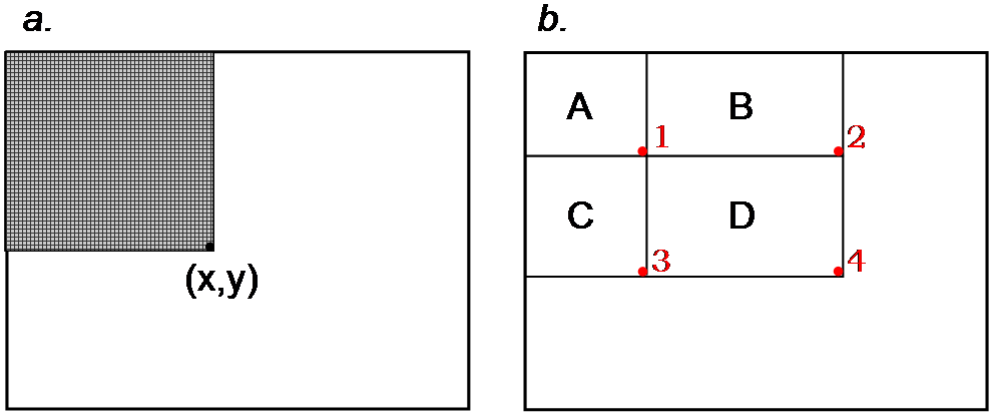

FIGURE A.2: The image of the left shows the shaded pixels and that x and y holds the value of these summed shaded pixels. The right-hand image details different rectangles and points to work out the pixel values in each rectangle or combination of rectangles. Image reproduced from [59].

contains the sum of the pixels above and to the left of (x, y), inclusive

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \tag{A.1}$$

where $ii(x, y)$ is the integral image and $i(x, y)$ is the original image. Using the following pair of recurrences

$$s(x, y) = s(x, y - 1) + i(x, y) \tag{A.2}$$

$$ii(x, y) = ii(x - 1, y) + s(x, y) \tag{A.3}$$

where $s(x, y)$ is the cumulative row sum, $s(x, -1) = 0$, and $ii(-1, y) = 0$ the integral image can be computed in one pass over the original image. Fig. A.2 a. shows shaded pixels to illustrate the SAT at x and y is the sum of the shaded pixels. Fig. A.2 b. shows how to obtain a summed value for another rectangle not using the top left of the image. To find the sum of pixel values in D can be described as: $D = A + B + C + D - (A + B) - (A + C) + A$. The numbered points 1-4 in Fig. A.2 *b.*, can be detailed as $x1, y1, x2$ and so on. With an Integral Image, you can find the sum of pixel values for any rectangle within the original image with this formula: $(x4, y4) - (x2, y2) - (x3, y3) + (x1, y1)$.

The next step is select which Haar features should be used and to set the thresholds mentioned previously. Selection of features is required because after all features have been calculated, there is around 160,000 features representing the image. Adaboost constructs a strong classifier as a linear combination of weak

classifiers, removing irrelevant features. Using only weak classifiers error rates for classification were as high as 0.5. A weak classifier is chosen if it can at least perform at chance accuracy.

The final step is to combine a series of Adaboost classifiers into a cascade of classifiers. A subregion of the image begins at the first classifier in a long chain of classifiers. If the image subregion passes classification through all of the cascade, then it is classified as a face. If it fails at any point in the cascade, it immediately is discarded as a true negative. The filters are also trained to pass an image subregion if it has also passed the previous filters.

Appendix B

Micro-Movement Detection Prototype

To address the limitations of only identifying and highlighting regions of the face, a micro-movement detection prototype is created. The prototype allows a user to input a movement and baseline video, and then it can be processed with the method developed in Chapter 7. The left hand side of Fig. B.1 shows the resulting micro-movement video with the cropped face and highlighted regions. The right side shows the feature difference graph.

As an initial prototype, the system follows the method exactly, however it takes around 10 minutes to process a video as the algorithm was not designed to be suitable for real-time applications. However, as the software was created in Matlab, further development in C++ with better optimisation would greatly increase the performance. Also, extra features such as video controls and parameter choosing would improve how a user would interact with the system.

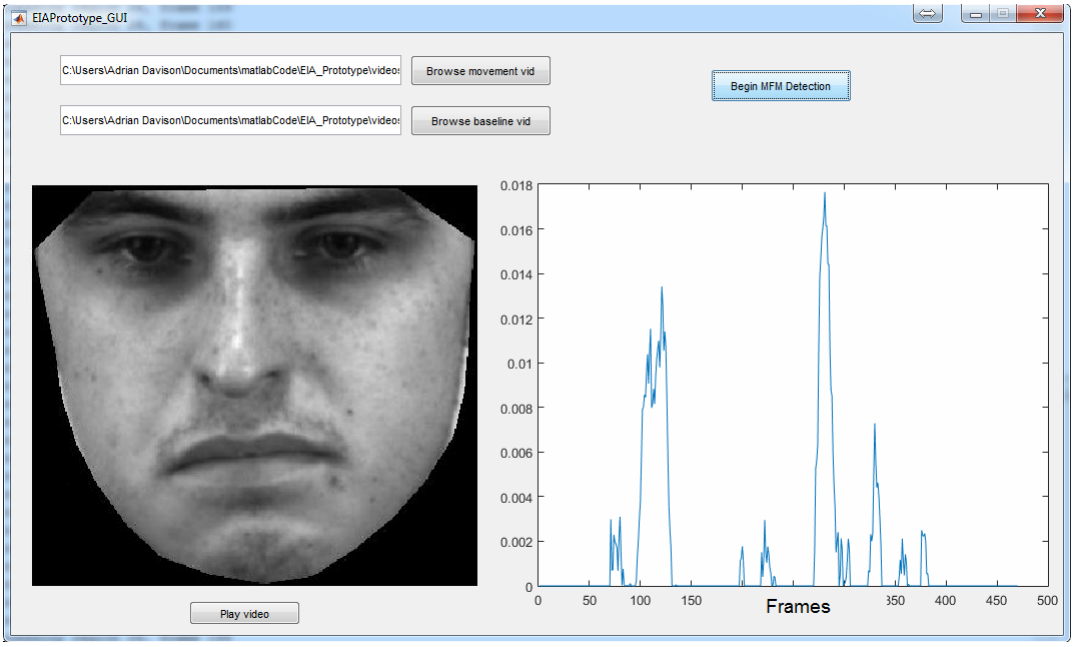

FIGURE B.1: A screenshot of the first prototype of the micro-movement detection system using FACS-based region analysis. The left side of the window is where the user can watch the micro-movement, and the feature difference values are on the right.

Bibliography

[1] Irfan A Essa and Alex P Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):757–763, Jul 1997. ISSN 0162-8828. doi: 10.1109/34.598232.

[2] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36:259–275, 2003.

[3] Jeffrey F. Cohn, Jing Xiao, Tsuyoshi Moriyama, Zara Ambadar, and Takeo Kanade. Automatic recognition of eye blinking in spontaneously occurring behavior. *Behavior Research Methods, Instruments, & Computers*, 35(3): 420–428, 2003. ISSN 0743-3808. doi: 10.3758/BF03195519. URL http://dx.doi.org/10.3758/BF03195519.

[4] Choon-Ching Ng, Moi Hoon Yap, Nicholas Costen, and Baihua Li. Automatic wrinkle detection using hybrid hessian filter. In *Computer Vision–ACCV 2014*, pages 609–622. Springer, 2015.

[5] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*, pages 16–21, 2009. URL http://digital-library.theiet.org/content/conferences/10.1049/ic.2009.0244.

[6] Matthew Shreve, Sridhar Godavarthy, Dmitry Goldgof, and Sudeep Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 51–56, 2011. doi: 10.1109/FG.2011.5771451.

[7] Sze-Teng Liong, John See, Raphael C-W Phan, Anh Cat Le Ngo, Yee-Hui Oh, and KokSheik Wong. Subtle expression recognition using optical strain weighted features. In *Computer Vision-ACCV 2014 Workshops*, pages 644–657. Springer, 2014.

[8] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, G. Zhao, and X.L. Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *Affective Computing, IEEE Transactions on*, PP(99): 1–1, 2015. ISSN 1949-3045. doi: 10.1109/TAFFC.2015.2485205.

[9] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. Reading hidden emotions: Spontaneous micro-expression spotting and recognition. *arXiv preprint arXiv:1511.00423*, 2015.

[10] Paul Ekman. *Emotions Revealed: Understanding Faces and Feelings*. Phoenix, 2004. ISBN 9780753817650.

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL http://dx.doi.org/10.1023/A%3A1022627411411.

[13] Mark G Frank, Carl J Maccario, and Venugopal l Govindaraju. Behavior and security. In *Protecting airline passengers in the age of terrorism*. Greenwood Pub. Group, 2009.

[14] (TSA) Transportation Security Administration. The truth behind the title: Behavior detection officer. blog.tsa.gov/2008/02/truth-behind-title-behavior-detection.html, 2008.

[15] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *10th IEEE International Conference on automatic Face and Gesture Recognition*, 2013.

[16] Wen-Jing Yan, Qi Wu, Yong-Jin Liu, Su-Jing Wang, and Xiaolan Fu. Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces. In *IEEE conference on automatic face and gesture recognition*, 2013.

[17] Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE*, 9(1):e86041, 01 2014. doi: http://dx.doi.org/10.1371%2Fjournal.pone.0086041.

[18] Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. Norton, 2001.

[19] Maureen O'Sullivan, Mark G Frank, Carolyn M Hurley, and Jaspreet Tiwana. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):530, 2009.

[20] Mark Frank, Malgorzata Herbasz, Kang Sinuk, Amy Marie Keller, Anastacia Kurylo, and Courtney Nolan. I see how you feel: Training laypeople and professionals to recognize fleeting emotions. In *International Communication Association*, 2009.

[21] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De La Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7, Sept 2009. doi: 10.1109/ACII.2009.5349358.

[22] Hanns C Hopf, Wibke Muller-Forell, and Nikolai J Hopf. Localization of emotional and volitional facial paresis. *Neurology*, 42(10):1918–1918, 1992.

[23] David Matsumoto, Seung Hee Yoo, and Sanae Nakagawa. Culture, emotion regulation, and adjustment. *Journal of personality and social psychology*, 94 (6):925, 2008.

[24] Tammy English and Oliver P John. Understanding the social effects of emotion regulation: the mediating role of authenticity for individual differences in suppression. *Emotion (Washington, D.C.)*, 13(2):314–329, April 2013. ISSN 1528-3542. doi: 10.1037/a0029847. URL http://dx.doi.org/10.1037/a0029847.

[25] Rossana B Queiroz, Soraia R Musse, and Norman I Badler. Investigating macroexpressions and microexpressions in computer graphics animated faces. *PRESENCE: Teleoperators and Virtual Environments*, 23(2):191–208, 2014.

[26] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[27] William E Rinn. The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological bulletin*, 95(1):52, 1984.

[28] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6: 169–200, 1992.

[29] Charles Darwin, Paul Ekman, and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. (Original work published in 1872).

[30] Takeo Kanade, Jeffrey F Cohn, and Ying Li Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53, 2000. doi: 10.1109/AFGR.2000.840611.

[31] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101, 2010. doi: 10.1109/CVPRW.2010.5543262.

[32] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5 pp.–, 2005. doi: 10.1109/ICME.2005.1521424.

[33] Henry Gray. *Anatomy of the human body*. Lea & Febiger, 1918.

[34] David Matsumoto, Dacher Keltner, Michelle N Shiota, MAUREEN OSullivan, and Mark Frank. *Facial expressions of emotion*, volume 3, chapter 13, pages 211–234. Guilford Publications New York, 2008.

[35] Guillermo O Paradiso, Danny I Cunic, Carolyn A Gunraj, and Robert Chen. Representation of facial muscles in human motor cortex. *The Journal of physiology*, 567(1):323–336, 2005.

[36] Paul Read and Mark-Paul Meyer. *Restoration of motion picture film*. Butterworth-Heinemann series in conservation and museology. Butterworth-Heinemann, 2000. ISBN 9780750627931. URL http://books.google.co.uk/books?id=OKZzxUV33zUC.

[37] James A Russell and Jose Miguel Fernández-Dols. *The psychology of facial expression*. Cambridge university press, 1997.

[38] Silvan Tomkins. *Affect Theory*, chapter 7, pages 163 – 195. Psychology Press, 1984.

[39] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[40] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.

[41] Paul Ekman. Lie catching and microexpressions. In Clancy W. Martin, editor, *The Philosophy of Deception*, pages 118–133. Oxford University Press, 2009.

[42] Xun-Bing Shen, Qi Wu, and Xiao-Lan Fu. Effects of the duration of expressions on the recognition of microexpressions. *Journal of Zhejiang University SCIENCE B*, 13(3):221–230, 2012. ISSN 1673-1581. doi: http://dx.doi.org/10.1631/jzus.B1100063.

[43] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37(4):217–230, 2013. ISSN 0191-5886. doi: http://dx.doi.org/10.1007/s10919-013-0159-8.

[44] Paul Ekman and Erika L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Series in Affective Science. Oxford University Press, 2005. ISBN 9780199792726.

[45] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[46] David Matsumoto and Hyi Sung Hwang. Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, 35:181–191, 2011.

[47] Emily A Butler, Tiane L Lee, and James J Gross. Emotion regulation and culture: are the social consequences of emotion suppression culture-specific? *Emotion (Washington, D.C.)*, 7(1):30–48, February 2007. ISSN 1528-3542. doi: 10.1037/1528-3542.7.1.30. URL http://dx.doi.org/10.1037/1528-3542.7.1.30.

[48] Pamela M. Cole, Margaret K. Michel, and Laureen O'Donnell Teti. The development of emotion regulation and dysregulation: A clinical perspective. *Monographs of the Society for Research in Child Development*, 59(2/3): pp. 73–100, 1994. ISSN 0037976X. URL http://www.jstor.org/stable/1166139.

[49] Elizabeth Davis, Ellen Greenberger, Susan Charles, Chuansheng Chen, Libo Zhao, and Qi Dong. Emotion experience and regulation in china and the united states: How do culture and gender shape emotion responding? *International Journal of Psychology*, 47(3):230–239, 2012. doi: 10.1080/00207594.2011.626043. URL http://dx.doi.org/10.1080/00207594.2011.626043.

[50] James J Gross. Emotion regulation: taking stock and moving forward. *Emotion (Washington, D.C.)*, 13(3):359–365, June 2013. ISSN 1528-3542. doi: 10.1037/a0032135. URL http://dx.doi.org/10.1037/a0032135.

[51] Amanda Sheffield Morris, Jennifer S. Silk, Laurence Steinberg, Sonya S. Myers, and Lara Rachel Robinson. The role of the family context in the development of emotion regulation. *Social Development*, 16(2):361–388, 2007. ISSN 1467-9507. doi: 10.1111/j.1467-9507.2007.00389.x. URL http://dx.doi.org/10.1111/j.1467-9507.2007.00389.x.

[52] Seung Hee Yoo, David Matsumoto, and Jeffrey A. LeRoux. The influence of emotion recognition and emotion regulation on intercultural adjustment. *International Journal of Intercultural Relations*, 30(3):345 – 363, 2006. ISSN 0147-1767. doi: http://dx.doi.org/10.1016/j.ijintrel.2005.08.006. URL http://www.sciencedirect.com/science/article/pii/S0147176705001136.

[53] Michael Biehl, David Matsumoto, Paul Ekman, Valerie Hearn, Karl Heider, Tsutomu Kudoh, and Veronica Ton. Matsumoto and ekman's japanese and caucasian facial expressions of emotion (jacfee): Reliability data and cross-national differences. *Journal of Nonverbal Behavior*, 21(1):3–21, 1997. ISSN 1573-3653. doi: 10.1023/A:1024902500935. URL http://dx.doi.org/10.1023/A:1024902500935.

[54] Paul Ekman and Richard J Davidson. Voluntary smiling changes regional brain activity. *Psychological Science*, 4(5):342–345, 1993.

[55] Tara L Kraft and Sarah D Pressman. Grin and bear it the influence of manipulated facial expression on the stress response. *Psychological science*, 23(11):1372–1378, 2012.

[56] Ying-Li Tian, Takeo Kanade, and Jeffrey F. Cohn. Facial expression analysis. In *Handbook of Face Recognition*, pages 247–275. Springer New York, 2011. ISBN 978-0-387-40595-7. doi: 10.1007/0-387-27257-7_12. URL http://dx.doi.org/10.1007/0-387-27257-7_12.

[57] Jeffrey F. Cohn. Foundations of human computing: Facial expression and emotion. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, ICMI '06, pages 233–238, New York, NY, USA, 2006. ACM. ISBN 1-59593-541-X. doi: 10.1145/1180995.1181043. URL http://doi.acm.org/10.1145/1180995.1181043.

[58] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511–I–518 vol.1, 2001. doi: 10.1109/CVPR.2001.990517.

[59] Rainer Lienhart, Er Kuranov, and Vadim Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *In DAGM 25th Pattern Recognition Symposium*, pages 297–304, 2003.

[60] Alex Pentland, Baback Moghaddam, and Thad Starner. View-based and modular eigenspaces for face recognition. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pages 84–91, Jun 1994. doi: 10.1109/CVPR.1994.323814.

[61] Stan Z. Li and L. Gu. Real-time multi-view face detection, tracking, pose estimation, alignment, and recognition. In *Computer Vision and Pattern Recognition*, 2001.

[62] Gary B Huang, Marwan Mattar, Honglak Lee, and Erik G Learned-Miller. Learning to align from scratch. In *In Neural Information Processing Systems*, 2012.

[63] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, 2001.

[64] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *International Conference on Computer Vision (ICCV)*, pages 1449–1456, nov. 2011. doi: 10.1109/ICCV.2011.6126401.

[65] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[66] Shiguang Shan, Yizheng Chang, Wen Gao, Bo Cao, and Peng Yang. Curse of mis-alignment in face recognition: problem and a novel mis-alignment learning solution. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 314–320, 2004. doi: 10.1109/AFGR.2004.1301550.

[67] Jianke Zhu, Luc Van Gool, and Steven C.H. Hoi. Unsupervised face alignment by robust nonrigid mapping. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1265–1272, 2009. doi: 10.1109/ICCV.2009.5459325.

[68] Yongqiang Li, Shangfei Wang, Yongping Zhao, and Qiang Ji. Simultaneous facial feature tracking and facial expression recognition. *Image Processing, IEEE Transactions on*, 22(7):2559–2573, July 2013. ISSN 1057-7149. doi: 10.1109/TIP.2013.2253477.

[69] Gwen Littlewort, Ian Fasel, Marian Stewart Bartlett, and Javier R. Movellan. Fully automatic coding of basic expressions from video. Technical report, Tech. rep.(2002) U of Calif., S.Diego, INC MPLab, 2002.

[70] Marian Stewart Bartlett, Gwen C Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, and Javier R Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[71] Ru-facs-1 database, 2004. URL http://mplab.ucsd.edu/grants/project1/research/rufacs1-dataset.html.

[72] John G. Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(7):1169–1179, Jul 1988. ISSN 0096-3518. doi: 10.1109/29.1644.

[73] Xijian Fan and Tardi Tjahjadi. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition*, 48(11):3407 – 3416, 2015. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2015.04.025. URL http://www.sciencedirect.com/science/article/pii/S0031320315001648.

[74] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 401–408, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-733-9. doi: 10.1145/1282280.1282340. URL http://doi.acm.org/10.1145/1282280.1282340.

[75] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, Feb 2001. ISSN 0162-8828. doi: 10.1109/34.908962.

[76] Zhen Wen and Thomas S Huang. Capturing subtle facial motions in 3d face tracking. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1343–1350 vol.2, Oct 2003. doi: 10.1109/ICCV.2003.1238646.

[77] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 229–234, May 2002. doi: 10.1109/AFGR.2002.1004159.

[78] Paul Ekman, Joseph Hager, C Methvin, and William Irwin. Ekmanhager facial action exemplars. Human Interaction Laboratory, University of California, San Francisco.

[79] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Transactions on Cybernetics*, 44(2):161–174, Feb 2014. ISSN 2168-2267. doi: 10.1109/TCYB.2013.2249063.

[80] Karan Sikka, Tingfan Wu, Josh Susskind, and Marian Bartlett. Exploring bag of words architectures in the facial expression domain. In *Proceedings of the 12th International Conference on Computer Vision - Volume 2*, ECCV'12, pages 250–259, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-33867-0. doi: 10.1007/978-3-642-33868-7_25. URL http://dx.doi.org/10.1007/978-3-642-33868-7_25.

[81] Michael Lyons, Shota Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205, 1998. doi: 10.1109/AFGR.1998.670949.

[82] O. Rudovic, M. Pantic, and I. Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1357–1369, June 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.233.

[83] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, Jan 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.20.

[84] Z. Wang, S. Wang, Y. Zhu, and Q. Ji. Bias analyses of spontaneous facial expression database. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2926–2929, Nov 2012.

[85] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1): 32–41, Jan 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.26.

[86] Shaohua Wan and J.K. Aggarwal. Spontaneous facial expression recognition: A robust metric learning approach. *Pattern Recognition*, 47(5):1859 – 1868, 2014. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2013.11.025. URL http://www.sciencedirect.com/science/article/pii/S0031320313005116.

[87] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, May 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.90.

[88] Naomi S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. doi: 10.1080/00031305.1992.10475879. URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879.

[89] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 37(6):1554–1563, 12 1966. doi: 10.1214/aoms/1177699147. URL http://dx.doi.org/10.1214/aoms/1177699147.

[90] Ying-Li Tian, Takeo Kanade, and Jeffrey Cohn. Eye-state action unit detection by gabor wavelets. In *Proceedings of the Third International Conference on Advances in Multimodal Interfaces (ICMI 2000)*, October 2000.

[91] Peng Yang, Qingshan Liu, and D.N. Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1018–1025, Sept 2009. doi: 10.1109/ICCV.2009.5459371.

[92] Jeffrey Cohn, Takeo Kanade, Tsuyoshi Moriyama, Zara Ambadar, Jing Xiao, Jiang Gao, and Hiroki Imamura. A comparative study of alternative facs coding algorithms. Technical Report CMU-RI-TR-02-06, Robotics Institute, Pittsburgh, PA, November 2001.

[93] Matthew Shreve, Jesse Brizzi, Sergiy Fefilatyev, Timur Luguev, Dmitry Goldgof, and Sudeep Sarkar. Automatic expression spotting in videos. *Image and Vision Computing*, 32(8):476 – 486, 2014. ISSN 0262-8856. doi: http://dx.doi.org/10.1016/j.imavis.2014.04.010.

[94] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1110.

[95] Zhaoyu Lu, Ziqi Luo, Huicheng Zheng, Jikai Chen, and Weihong Li. A delaunay-based temporal coding model for micro-expression recognition. In *Computer Vision-ACCV 2014 Workshops*, pages 698–711. Springer, 2014.

[96] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, and Xiaolan Fu. Micro-expression recognition using dynamic textures on tensor independent color space. In *ICPR*, 2014.

[97] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces. *Image Processing, IEEE Transactions on*, PP(99): 1–1, 2015. ISSN 1057-7149. doi: 10.1109/TIP.2015.2496314.

[98] Senya Polikovsky, Yoshinari Kameda, and Ohta Yuichi. Facial micro-expression detection in hi-speed video based on facial action coding system (facs). *IEICE transactions on information and systems*, 96(1):81–92, 2013.

[99] Qi Wu, Xunbing Shen, and Xiaolan Fu. The machine knows what you are hiding: an automatic micro-expression recognition system. In *Affective Computing and Intelligent Interaction*, pages 152–162. Springer, 2011.

[100] Paul Ekman. Mett. *Micro Expression Training Tool*, 2003.

[101] Mark G. Frank and Paul Ekman. The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, 72:1429–1439, 1997.

[102] Michel Owayjan, Ahmad Kashour, Nancy Al Haddad, Mohamad Fadel, and Ghinwa Al Souki. The design and development of a lie detection system using facial micro-expressions. In *2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, pages 33–38, December 2012. doi: 10.1109/ICTEA.2012.6462897.

[103] Yale Song, Louis-Philippe Morency, and Randall Davis. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In

Proceedings of the 15th ACM on International conference on multimodal interaction, pages 237–244. ACM, 2013.

[104] Yanjun Guo, Yantao Tian, Xu Gao, and Xuange Zhang. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 3473–3479. IEEE, 2014.

[105] Su-Jing Wang, Hui-Ling Chen, Wen-Jing Yan, Yu-Hsin Chen, and Xiaolan Fu. Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural processing letters*, 39(1):25–43, 2014.

[106] Yandan Wang, John See, Raphael C.-W. Phan, and Yee-Hui Oh. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PLoS ONE*, 10(5):e0124674, 05 2015. doi: 10.1371/journal.pone.0124674. URL http://dx.doi.org/10.1371%2Fjournal.pone.0124674.

[107] Matthew Shreve, Sridhar Godavarthy, Vasant Manohar, Dmitry Goldgof, and Sudeep Sarkar. Towards macro- and micro-expression spotting in video using strain patterns. In *2009 Workshop on Applications of Computer Vision (WACV)*, pages 1 –6, dec. 2009. doi: 10.1109/WACV.2009.5403044.

[108] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–4, Sept 2009. doi: 10.1109/ACII.2009.5349466.

[109] Antti Moilanen, Guoying Zhao, and Matti Pietikainen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1722–1727, Aug 2014. doi: 10.1109/ICPR.2014.303.

[110] Paul Ekman, Maureen O'Sullivan, Wallace V. Friesen, and Klaus R. Scherer. Invited article: Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, 15(2):125–135, 1991. ISSN 0191-5886. doi: 10.1007/BF00998267. URL http://dx.doi.org/10.1007/BF00998267.

[111] Brian E Malone and Bella M DePaulo. *Interpersonal sensitivity: Theory and measurement*, chapter Measuring sensitivity to deception, pages 103–124. Psychology Press, 2001.

[112] Zhaoqiang Xia, Xiaoyi Feng, Jinye Peng, Xianlin Peng, and Guoying Zhao. Spontaneous micro-expression spotting via geometric deformation modeling. *Computer Vision and Image Understanding*, 2015. ISSN 1077-3142. doi: http://dx.doi.org/10.1016/j.cviu.2015.12.006. URL http://www.sciencedirect.com/science/article/pii/S1077314215002702.

[113] Devangini Patel, Guoying Zhao, and Matti Pietikäinen. Spatiotemporal integration of optical flow vectors for micro-expression detection. In *Advanced Concepts for Intelligent Vision Systems*, pages 369–380. Springer, 2015.

[114] RM McCabe. Best practice recommendation for the capture of mugshots version 2.0. Available at: http://biometrics.nist.gov/cs_links/standard/ansi_2010/archive/Best_Practice_Face_Pose_Value.pdf, 1997.

[115] Shazia Afzal and Peter Robinson. Natural affect datacollection & annotation in a learning context. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009.

[116] Gemma Warren, Elizabeth Schertler, and Peter Bull. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior*, 33: 59–69, 2009. ISSN 0191-5886. doi: 10.1007/s10919-008-0057-7. URL http://dx.doi.org/10.1007/s10919-008-0057-7.

[117] Alejandro Jaimes and Nicu Sebe. Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(12): 116 – 134, 2007. ISSN 1077-3142. doi: http://dx.doi.org/10.1016/j.cviu.2006.10.019. URL http://www.sciencedirect.com/science/article/pii/S1077314206002335. Special Issue on Vision for Human-Computer Interaction.

[118] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, 1(1):137–159,

March 2008. URL http://www.s2is.org/Issues/v1/n1/papers/paper9.pdf.

[119] Maureen O'Sullivan, Mark G. Frank, Carolyn M. Hurley, and Jaspreet Tiwana. Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):pp. 530–538, 542–543, 2009. ISSN 01477307. URL http://www.jstor.org/stable/40540290.

[120] Leonard Saxe. Lying: Thoughts of an applied social psychologist. *American Psychologist*, 46(4):409, 1991.

[121] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. ISSN 0022-0000. doi: http://dx.doi.org/10.1006/jcss.1997.1504. URL http://www.sciencedirect.com/science/article/pii/S002200009791504X.

[122] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 386–391, Dec 2013. doi: 10.1109/ICCVW.2013.58.

[123] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 397–403, Dec 2013. doi: 10.1109/ICCVW.2013.59.

[124] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[125] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012.

[126] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807 – 813, 2010. ISSN 0262-8856. doi: http://dx.doi.org/10.1016/j.imavis.2009.08.002. URL http://www.sciencedirect.com/science/article/pii/S0262885609001711. Best of Automatic Face and Gesture Recognition 2008.

[127] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 545–552, June 2011. doi: 10.1109/CVPR.2011.5995602.

[128] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886, 2012. doi: 10.1109/CVPR.2012.6248014.

[129] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, pages –, 2016. ISSN 0262-8856. doi: http://dx.doi.org/10.1016/j.imavis.2016.01.002. URL http://www.sciencedirect.com/science/article/pii/S0262885616000147.

[130] D Cristinacce and TF Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 3, pages 929–938. BMVA, 2006.

[131] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054 – 3067, 2008. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2008.01.024. URL http://www.sciencedirect.com/science/article/pii/S0031320308000630.

[132] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, June 2014. doi: 10.1109/CVPR.2014.239.

[133] Li Zhang, Kamlesh Mistry, Ming Jiang, Siew Chin Neoh, and Mohammed Alamgir Hossain. Adaptive facial point detection and emotion recognition for a humanoid robot. *Computer Vision and Image Understanding*, 140:93 – 114, 2015. ISSN 1077-3142. doi: http://dx.doi.org/10.1016/j.cviu.2015.07.007. URL http://www.sciencedirect.com/science/article/pii/S1077314215001605.

[134] Javier Orozco, Ognjen Rudovic, Jordi Gonzlez, and Maja Pantic. Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, 31(4):322 – 340, 2013. ISSN

0262-8856. doi: http://dx.doi.org/10.1016/j.imavis.2013.02.001. URL http://www.sciencedirect.com/science/article/pii/S0262885613000334.

[135] Timothy F Cootes and Panachit Kittipanya-ngam. Comparing variations on the active appearance model algorithm. In *BMVC*, pages 1–10, 2002.

[136] Timothy F Cootes, Cristopher J Taylor, et al. Statistical models of appearance for computer vision, 2004.

[137] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[138] Manuel Guizar-Sicairos, Samuel T. Thurman, and James R. Fienup. Efficient subpixel image registration algorithms. *Opt. Lett.*, 33(2):156–158, Jan 2008. doi: 10.1364/OL.33.000156. URL http://ol.osa.org/abstract.cfm?URI=ol-33-2-156.

[139] Kostadin Dabov, Alessandro Foi, and Karen Egiazarian. Video denoising by sparse 3d transform-domain collaborative filtering. In *Proc. 15th European Signal Processing Conference*, volume 1, page 7, 2007.

[140] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7): 971–987, Jul 2002. ISSN 0162-8828. doi: 10.1109/TPAMI.2002.1017623.

[141] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *BMVC*, volume 2, page 5, 2009.

[142] Piotr Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[143] Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical symposium east*, pages 319–331. International Society for Optics and Photonics, 1981.

[144] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981.

[145] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and Rene Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear

dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939, June 2009. doi: 10.1109/CVPR.2009.5206821.

[146] Jasper Uijlings, Ionut C. Duta, Enver Sangineto, and Nicu Sebe. Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. *International Journal of Multimedia Information Retrieval*, 4(1):33–44, 2014.

[147] Ofir Pele and Michael Werman. The quadratic-chi histogram distance family. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision ECCV 2010*, volume 6312 of *Lecture Notes in Computer Science*, pages 749–762. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15551-2.

[148] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037–2041, Dec 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.244.

[149] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2): 213–238, 2007.

[150] Tom O'Haver. Signal processing tools. https://terpconnect.umd.edu/~toh/spectrum/SignalProcessingTools.html.

[151] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[152] Adrian K. Davison, Moi Hoon Yap, Nicholas Costen, Kevin Tan, Cliff Lansley, and Daniel Leightley. Micro-facial movements: An investigation on spatio-temporal descriptors. In *ECCVW*, 2014.

[153] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[154] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[155] Bart M Haar Romeny. Gaussian derivatives. In *Front-End Vision and Multi-Scale Image Analysis*. Springer Netherlands, 2003. ISBN 978-1-4020-1503-8.

[156] LMJ Florack, BM Ter Haar Romeny, Jan J Koenderink, and Max A Viergever. General intensity transformations and differential invariants. *Journal of Mathematical Imaging and Vision*, 4(2):171–187, 1994.

[157] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157 vol.2, 1999. doi: 10.1109/ICCV.1999.790410.

[158] Philippe Cattin. Image restoration: Introduction to signal and image processing, April 2015.

[159] Toni Buades, Yifei Lou, Jean-Michel Morel, and Zhongwei Tang. A note on multi-image denoising. In *Local and Non-Local Approximation in Image Processing, 2009. LNLA 2009. International Workshop on*, pages 1–15, Aug 2009. doi: 10.1109/LNLA.2009.5278408.

[160] James C. Brailean, Richard P. Kleihorst, Serafim. Efstratiadis, Aggelos K. Katsaggelos, and Reginald L. Lagendijk. Noise reduction filters for dynamic image sequences: a review. *Proceedings of the IEEE*, 83(9):1272–1292, Sep 1995. ISSN 0018-9219. doi: 10.1109/5.406412.

[161] Matan Protter and Michael Elad. Image sequence denoising via sparse and redundant representations. *Image Processing, IEEE Transactions on*, 18(1): 27–35, Jan 2009. ISSN 1057-7149. doi: 10.1109/TIP.2008.2008065.

[162] Alessandra Mammucari, Calogero Caltagirone, Paul Ekman, Wallace V. Friesen, Guido Gainotti, Luigi Pizzamiglio, and Pierluigi Zoccolotti. Spontaneous facial expression of emotions in brain-damaged patients. *Cortex*, 24 (4):521–533, 1988.

[163] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9): 1627–1645, 2010.

[164] Megvii Inc. Face++ research toolkit. www.faceplusplus.com, December 2013.

[165] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[166] Adrian K. Davison, Moi Hoon Yap, and Cliff Lansley. Micro-facial movement detection using individualised baselines and histogram-based descriptors. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 1864–1869, Oct 2015. doi: 10.1109/SMC.2015.326.

[167] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on*, volume 5, pages 53–53, June 2003. doi: 10.1109/CVPRW.2003.10057.

[168] Jaewon Sung and Daijin Kim. Real-time facial expression recognition using {STAAM} and layered {GDA} classifier. *Image and Vision Computing*, 27(9):1313 – 1325, 2009. ISSN 0262-8856. doi: http://dx.doi.org/10.1016/j.imavis.2008.11.010. URL http://www.sciencedirect.com/science/article/pii/S0262885608002527.

[169] Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer. Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4): 966–979, Aug 2012. ISSN 1083-4419. doi: 10.1109/TSMCB.2012.2200675.